Handbook of Financial Time Series

Torben G. Andersen • Richard A. Davis
Jens-Peter Kreiß • Thomas Mikosch
Editors

Handbook of Financial Time Series

Springer

Torben G. Andersen
Department of Finance
Kellogg School of Management
Northwestern University
2001 Sheridan Road
Evanston, IL 60208
U.S.A.
t-andersen@kellogg.northwestern.edu

Richard A. Davis
Department of Statistics
Columbia University
1255 Amsterdam Avenue
New York, NY 10027
U.S.A.
rdavis@stat.columbia.edu

Jens-Peter Kreiß
Institut für Mathematische Stochastik
Technische Universität Braunschweig
Pockelsstrasse 14
38106 Braunschweig
Germany
j.kreiss@tu-bs.de

Thomas Mikosch
Department Mathematics
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark
mikosch@math.ku.dk

ISBN 978-3-540-71296-1 e-ISBN 978-3-540-71297-8

Library of Congress Control Number: 2008943984

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Foreword

The *Handbook of Financial Time Series*, edited by Andersen, Davis, Kreiss and Mikosch, is an impressive collection of survey articles by many of the leading contributors to the field. These articles are mostly very clearly written and present a sweep of the literature in a coherent pedagogical manner. The level of most of the contributions is mathematically sophisticated, and I imagine many of these chapters will find their way onto graduate reading lists in courses in financial economics and financial econometrics. In reading through these papers, I found many new insights and presentations even in areas that I know well.

The book is divided into five broad sections: GARCH-Modeling, Stochastic Volatility Modeling, Continuous Time Processes, Cointegration and Unit Roots, and Special Topics. These correspond generally to classes of stochastic processes that are applied in various finance contexts. However, there are other themes that cut across these classes. There are several papers that carefully articulate the probabilistic structure of these classes, while others are more focused on estimation. Still others derive properties of extremes for each class of processes, and evaluate persistence and the extent of long memory. Papers in many cases examine the stability of the process with tools to check for breaks and jumps. Finally there are applications to options, term structure, credit derivatives, risk management, microstructure models and other forecasting settings.

The GARCH family of models is nicely introduced by Teräsvirta and then the mathematical underpinning is elegantly and readably presented by Lindner with theorems on stationarity, ergodicity and tail behavior. In the same vein, Giraitis, Leipus and Surgailis examine the long memory properties of infinite order ARCH models with memory decay slower than GARCH, and Davis and Mikosch derive tail properties of GARCH models showing that they satisfy a power law and are in the maximum domain of attraction of the Fréchet distribution. The multivariate GARCH family is well surveyed by Silvennoinen and Teräsvirta. Linton and Čížek and Spokoiny, respectively, specify models which are non- or semi-parametric or which are only constant over intervals of homogeneity.

The section on Stochastic Volatility Modelling (SV) brings us up to date on the development of alternatives to GARCH style models. Davis and Mikosch in two chapters develop the somewhat easier underlying mathematical theory and tail properties of SV. They derive an important difference from GARCH models. While both stochastic volatility and GARCH processes exhibit volatility clustering, only the GARCH has clustering of extremes. Long memory is conveniently described by SV models in Hurvich and Soulier. Chib, Omori and Asai extend these analyses to multivariate systems although they do not envision very high dimensions. Estimation is covered in several chapters by Renault, Shephard and Andersen, and Jungbacker and Koopman.

The continuous time analysis begins with familiar Brownian motion processes and enhances them with jumps, dynamics, time deformation, correlation with returns and Lévy process innovations. Extreme value distributions are developed and estimation algorithms for discretely sampled processes are analyzed. Lindner discusses the idea of continuous time approximations to GARCH and SV models showing that the nature of the approximation must be carefully specified. The continuous time framework is then applied to several finance settings such as interest rate models by Björk, option pricing by Kallsen, and realized volatility by Andersen and Benzoni. The book then returns to analysis of first moments with surveys of discrete time models with unit roots, near unit roots, fractional unit roots and cointegration.

Finally, a remaining 13 chapters are collected in a section called Special Topics. These include very interesting chapters on copulas, non-parametric models, resampling methods, Markov switching models, structural break models and model selection. Patton and Sheppard examine univariate and multivariate volatility forecast comparisons. They show the advantages of a GLS correction, discuss multiple comparisons and economic loss functions. Bauwens and Hautsch survey a wide range of models for point processes that have been used in the finance literature to model arrival times of trades and quotes. The survey is well grounded in the statistical literature and the economics literature. Embrechts, Furrer and Kaufmann discuss different types of risk— market, credit, operational and insurance—and some of the leading approaches to estimation. Christoffersen applies the filtered historical simulation or FHS method to univariate and multivariate simulation based calculation of VaR, Expected Shortfall and active portfolio risks. Lando surveys the structural and reduced form approaches to modeling credit spreads. He focuses on CDS spreads and default dependence and gives a nice description of tests between contagion and factor structures in formulating dependence.

So make yourself a double cappuccino and relax in a comfortable chair, or adjust your headphones at 30,000 ft. over the Pacific, and dig in. There are treats in lots of different areas just waiting to be discovered.

New York, September 2008 *Robert Engle*

Contents

Part V Special Topics – Simulation Based Methods

List of Contributors

Yacine Aït-Sahalia
Princeton University and NBER, Bendheim Center for Finance, Princeton University, U.S.A..

Torben G. Andersen
Kellogg School of Management and NBER, Northwestern University, U.S.A. and CREATES, Aarhus, Denmark.

Elena Andreou
Department of Economics, University of Cyprus, Nicosia, Cyprus.

Manabu Asai
Faculty of Economics, Soka University, Tokyo, Japan.

Luc Bauwens
CORE, Université Catholique de Louvain, Belgium.

Luca Benzoni
Federal Reserve Bank of Chicago, U.S.A..

Thomas Björk
Department of Finance, Stockholm School of Economics, Sweden.

Peter J. Brockwell
Department of Statistics, Colorado State University, Fort Collins, U.S.A..

Ngai Hang Chan
Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong.

Willa W. Chen
Department of Statistics, Texas A&M University, College Station, U.S.A..

Siddhartha Chib
Olin Business School, Washington University in St. Louis, U.S.A..

Peter Christoffersen
Desautels Faculty of Management, McGill University, Quebec, Canada.

Pavel Čížek
Department of Econometrics & OR, Tilburg University, The Netherlands.

Richard A. Davis
Department of Statistics, Columbia University, New York, U.S.A..

Ernst Eberlein
Department of Mathematical Stochastics, University of Freiburg, Germany.

Paul Embrechts
Department of Mathematics, ETH Zürich, Switzerland.

Vicky Fasen
Zentrum Mathematik, Technische Universität München, Germany.

Christian Francq
University Lille III, EQUIPPE-GREMARS, France.

Jürgen Franke
Department of Mathematics, Universität Kaiserslautern, Germany.

Hansjörg Furrer
Swiss Life, Zürich, Switzerland.

Eric Ghysels
Department of Economics, University of North Carolina at Chapel Hill, U.S.A..

Liudas Giraitis
Department of Economics, Queen Mary University of London, United Kingdom.

Nikolaus Hautsch
Institute for Statistics and Econometrics, Humboldt–Universität zu Berlin, Germany.

Clifford M. Hurvich
Leonard N. Stern School of Business, New York University, U.S.A..

Michael Johannes
Graduate School of Business, Columbia University, New York, U.S.A..

Søren Johansen
Department of Applied Mathematics and Statistics, University of Copenhagen, Denmark.

Borus Jungbacker
Department of Econometrics, Vrije Universiteit Amsterdam, The Netherlands.

Jan Kallsen
Mathematisches Seminar, Christian-Albrechts-Universität zu Kiel, Germany.

Roger Kaufmann
AXA Winterthur, Winterthur, Switzerland.

Siem Jan Koopman
Department of Econometrics, Vrije Universiteit Amsterdam, The Netherlands.

Jens-Peter Kreiss
Institut für Mathematische Stochastik, Technische Universität Braunschweig, Germany.

David Lando
Copenhagen Business School, Department of Finance, Denmark.

Theis Lange
Department of Economics, University of Copenhagen, Denmark.

Hannes Leeb
Department of Statistics, Yale University, U.S.A..

Remigijus Leipus
Vilnius University and Institute of Mathematics and Informatics, Vilnius, Lithuania.

Alexander M. Lindner
Technische Universität Braunschweig, Institut für Mathematische Stochastik, Germany.

Oliver B. Linton
Department of Economics, London School of Economics and Political Science, United Kingdom.

Ross A. Maller
School of Finance & Applied Statistics and Centre for Mathematics & its Applications, Australian National University, Canberra, Australia.

Enno Mammen
Abteilung Volkswirtschaftslehre, Universität Mannheim, Germany.

Thomas Mikosch
Laboratory of Actuarial Mathematics, University of Copenhagen, Denmark.

Gernot Müller
Zentrum Mathematik, Technische Universität München, Germany.

Per A. Mykland
Department of Statistics, The University of Chicago, U.S.A..

Yasuhiro Omori
Faculty of Economics, University of Tokyo, Japan.

Efstathios Paparoditis
Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus.

Andrew J. Patton
Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, United Kingdom.

Peter C. B. Phillips
Cowles Foundation for Research in Economics, Yale University, U.S.A.; University of Auckland; University of York; and Singapore Management University.

Benedikt M. Pötscher
Department of Statistics, University of Vienna, Austria.

Dimitris N. Politis
Department of Mathematics, University of California, San Diego, U.S.A..

Nicholas Polson
Graduate School of Business, University of Chicago, U.S.A..

Anders Rahbek
Department of Economics, University of Copenhagen, Denmark.

Eric Renault
Department of Economics, University of North Carolina, Chapel Hill, U.S.A..

Neil Shephard
Oxford-Man Institute and Department of Economics, University of Oxford, United Kingdom.

Kevin Sheppard
Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, United Kingdom.

Annastiina Silvennoinen
School of Finance and Economics, University of Technology Sydney, Australia.

Michael Sørensen
Department of Mathematical Sciences, University of Copenhagen, Denmark.

Philippe Soulier
Department of Mathematics, University Paris X, France.

Vladimir Spokoiny
Weierstrass–Institut, Berlin, Germany.

Donatas Surgailis
Vilnius University and Institute of Mathematics and Informatics, Vilnius, Lithuania.

Alex Szimayer
Fraunhofer-Institut für Techno-und Wirtschaftsmathematik, Kaiserslautern, Germany.

Timo Teräsvirta
CREATES, School of Economics and Management, University of Aarhus, Denmark and Department of Economic Statistics, Stockholm School of Economics, Sweden.

Jun Yu
School of Economics, Singapore Management University, Singapore.

Jean-Michel Zakoïan
University Lille III, EQUIPPE-GREMARS, and CREST, France.

Eric Zivot
Department of Economics, University of Washington, Seattle, U.S.A..

An Introduction to Univariate GARCH Models

Timo Teräsvirta

Abstract This paper contains a survey of univariate models of conditional heteroskedasticity. The classical ARCH model is mentioned, and various extensions of the standard Generalized ARCH model are highlighted. This includes the Exponential GARCH model. Stochastic volatility models remain outside this review.

1 Introduction

Financial economists are concerned with modelling volatility in asset returns. This is important as volatility is considered a measure of risk, and investors want a premium for investing in risky assets. Banks and other financial institutions apply so-called value-at-risk models to assess their risks. Modelling and forecasting volatility or, in other words, the covariance structure of asset returns, is therefore important.

The fact that volatility in returns fluctuates over time has been known for a long time. Originally, the emphasis was on another aspect of return series: their marginal distributions were found to be leptokurtic. Returns were modelled as independent and identically distributed over time. In a classic work, Mandelbrot (1963) and Mandelbrot and Taylor (1967) applied so-called stable Paretian distributions to characterize the distribution of returns. Rachev and Mittnik (2000) contains an informative discussion of stable Paretian distributions and their use in finance and econometrics.

Observations in return series of financial assets observed at weekly and higher frequencies are in fact not independent. While observations in these

Timo Teräsvirta
CREATES, School of Economics and Management, University of Aarhus, DK-8000 Aarhus C, and Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, e-mail: tterasvirta@econ.au.dk

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_1,

series are uncorrelated or nearly uncorrelated, the series contain higher-order dependence. Models of Autoregressive Conditional Heteroskedasticity (ARCH) form the most popular way of parameterizing this dependence. There are several articles in this *Handbook* devoted to different aspects of ARCH models. This article provides an overview of different parameterizations of these models and thus serves as an introduction to autoregressive conditional heteroskedasticity. The article is organized as follows. Section 2 introduces the classic ARCH model. Its generalization, the Generalized ARCH (GARCH) model is presented in Section 3. This section also describes a number of extensions to the standard GARCH models. Section 4 considers the Exponential GARCH model whose structure is rather different from that of the standard GARCH model, and Section 5 discusses ways of comparing EGARCH models with GARCH ones. Suggestions for further reading can be found at the end.

2 The ARCH Model

The autoregressive conditional heteroskedasticity (ARCH) model is the first model of conditional heteroskedasticity. According to Engle (2004), the original idea was to find a model that could assess the validity of the conjecture of Friedman (1977) that the unpredictability of inflation was a primary cause of business cycles. Uncertainty due to this unpredictability would affect the investors' behaviour. Pursuing this idea required a model in which this uncertainty could change over time. Engle (1982) applied his resulting ARCH model to parameterizing conditional heteroskedasticity in a wage-price equation for the United Kingdom. Let ε_t be a random variable that has a mean and a variance conditionally on the information set $\mathcal{F}_{t-1}$ (the σ-field generated by ε_{t-j}, $j \geq 1$). The ARCH model of ε_t has the following properties. First, $\mathsf{E}\{\varepsilon_t|\mathcal{F}_{t-1}\} = 0$ and, second, the conditional variance $h_t = \mathsf{E}\{\varepsilon_t^2|\mathcal{F}_{t-1}\}$ is a nontrivial positive-valued parametric function of $\mathcal{F}_{t-1}$. The sequence $\{\varepsilon_t\}$ may be observed directly, or it may be an error or innovation sequence of an econometric model. In the latter case,

$$\varepsilon_t = y_t - \mu_t(y_t) \tag{1}$$

where y_t is an observable random variable and $\mu_t(y_t) = \mathsf{E}\{y_t|\mathcal{F}_{t-1}\}$, the conditional mean of y_t given $\mathcal{F}_{t-1}$. Engle's (1982) application was of this type. In what follows, the focus will be on parametric forms of h_t, and for simplicity it is assumed that $\mu_t(y_t) = 0$.

Engle assumed that ε_t can be decomposed as follows:

$$\varepsilon_t = z_t h_t^{1/2} \tag{2}$$

where $\{z_t\}$ is a sequence of independent, identically distributed (iid) random variables with zero mean and unit variance. This implies $\varepsilon_t|\mathcal{F}_{t-1} \sim D(0, h_t)$ where D stands for the distribution (typically assumed to be a normal or a leptokurtic one). The following conditional variance defines an ARCH model of order q:

$$h_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 \tag{3}$$

where $\alpha_0 > 0, \alpha_j \geq 0, j = 1, \ldots, q-1$, and $\alpha_q > 0$. The parameter restrictions in (3) form a necessary and sufficient condition for positivity of the conditional variance. Suppose the unconditional variance $\mathsf{E}\varepsilon_t^2 = \sigma^2 < \infty$. The definition of ε_t through the decomposition (2) involving z_t then guarantees the white noise property of the sequence $\{\varepsilon_t\}$, since $\{z_t\}$ is a sequence of iid variables. Although the application in Engle (1982) was not a financial one, Engle and others soon realized the potential of the ARCH model in financial applications that required forecasting volatility.

The ARCH model and its generalizations are thus applied to modelling, among other things, interest rates, exchange rates and stock and stock index returns. Bollerslev et al. (1992) already listed a variety of applications in their survey of these models. Forecasting volatility of these series is different from forecasting the conditional mean of a process because volatility, the object to be forecast, is not observed. The question then is how volatility should be measured. Using ε_t^2 is an obvious but not necessarily a very good solution if data of higher frequency are available; see Andersen and Bollerslev (1998) and Andersen and Benzoni (2008) for discussion.

3 The Generalized ARCH Model

3.1 Why Generalized ARCH?

In applications, the ARCH model has been replaced by the so-called generalized ARCH (GARCH) model that Bollerslev (1986) and Taylor (1986) proposed independently of each other. In this model, the conditional variance is also a linear function of its own lags and has the form

$$h_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}. \tag{4}$$

The conditional variance defined by (4) has the property that the unconditional autocorrelation function of ε_t^2, if it exists, can decay slowly, albeit still exponentially. For the ARCH family, the decay rate is too rapid compared to what is typically observed in financial time series, unless the maximum

lag q in (3) is long. As (4) is a more parsimonious model of the conditional variance than a high-order ARCH model, most users prefer it to the simpler ARCH alternative.

The overwhelmingly most popular GARCH model in applications has been the GARCH(1,1) model, that is, $p = q = 1$ in (4). A sufficient condition for the conditional variance to be positive is $\alpha_0 > 0$, $\alpha_j \geq 0$, $j = 1, \ldots, q$; $\beta_j \geq 0$, $j = 1, \ldots, p$. The necessary and sufficient conditions for positivity of the conditional variance in higher-order GARCH models are more complicated than the sufficient conditions just mentioned and have been given in Nelson and Cao (1992). The GARCH(2,2) case has been studied in detail by He and Teräsvirta (1999). Note that for the GARCH model to be identified if at least one $\beta_j > 0$ (the model is a genuine GARCH model) one has to require that also at least one $\alpha_j > 0$. If $\alpha_1 = \ldots = \alpha_q = 0$, the conditional and unconditional variances of ε_t are equal and $\beta_1, \ldots, \beta_p$ are unidentified nuisance parameters. The GARCH(p,q) process is weakly stationary if and only if $\sum_{j=1}^{q} \alpha_j + \sum_{j=1}^{p} \beta_j < 1$.

The stationary GARCH model has been slightly simplified by 'variance targeting', see Engle and Mezrich (1996). This implies replacing the intercept α_0 in (4) by $(1 - \sum_{j=1}^{q} \alpha_j - \sum_{j=1}^{p} \beta_j)\sigma^2$ where $\sigma^2 = \mathsf{E}\varepsilon_t^2$. The estimate $\widehat{\sigma}^2 = T^{-1} \sum_{t=1}^{T} \varepsilon_t^2$ is substituted for σ^2 before estimating the other parameters. As a result, the conditional variance converges towards the 'long-run' unconditional variance, and the model contains one parameter less than the standard GARCH(p,q) model.

It may be pointed out that the GARCH model is a special case of an infinite-order (ARCH(∞)) model (2) with

$$h_t = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j \varepsilon_{t-j}^2. \tag{5}$$

The ARCH(∞) representation is useful in considering properties of ARCH and GARCH models such as the existence of moments and long memory; see Giraitis et al. (2000). The moment structure of GARCH models is considered in detail in Lindner (2008).

3.2 Families of univariate GARCH models

Since its introduction the GARCH model has been generalized and extended in various directions. This has been done to increase the flexibility of the original model. For example, the original GARCH specification assumes the response of the variance to a shock to be independent of the sign of the shock and just be a function of the size of the shock. Several extensions of the GARCH model aim at accommodating the asymmetry in the response.

These include the GJR-GARCH model of Glosten et al. (1993), the asymmetric GARCH models of Engle and Ng (1993) and the quadratic GARCH of Sentana (1995). The GJR-GARCH model has the form (2), where

$$h_t = \alpha_0 + \sum_{j=1}^{q} \{\alpha_j + \delta_j I(\varepsilon_{t-j} > 0)\} \varepsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}. \tag{6}$$

In 6), $I(\varepsilon_{t-j} > 0)$ is an indicator function obtaining value one when the argument is true and zero otherwise.

In the asymmetric models of both Engle and Ng, and Sentana, the centre of symmetry of the response to shocks is shifted away from zero. For example,

$$h_t = \alpha_0 + \alpha_1 (\varepsilon_{t-1} - \gamma)^2 + \beta_1 h_{t-1} \tag{7}$$

with $\gamma \neq 0$ in Engle and Ng (1993). The conditional variance in Sentana's Quadratic ARCH (QARCH) model (the model is presented in the ARCH form) is defined as follows:

$$h_t = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\varepsilon}_{t-1}' \boldsymbol{A} \boldsymbol{\varepsilon}_{t-1} \tag{8}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \ldots, \varepsilon_{t-q+1})'$ is a $q \times 1$ vector, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'$ is a $q \times 1$ parameter vector and $\boldsymbol{A}$ a $q \times q$ symmetric parameter matrix. In (8), not only squares of ε_{t-i} but also cross-products $\varepsilon_{t-i}\varepsilon_{t-j}$, $i \neq j$, contribute to the conditional variance. When $\boldsymbol{\alpha} \neq \mathbf{0}$, the QARCH generates asymmetric responses. The ARCH equivalent of (7) is a special case of Sentana's model. Constraints on parameters required for positivity of h_t in (8) become clear by rewriting (8) as follows:

$$h_t = \begin{bmatrix} \boldsymbol{\varepsilon}_{t-1} & 1 \end{bmatrix}' \begin{bmatrix} \boldsymbol{A} & \boldsymbol{\alpha}/2 \\ \boldsymbol{\alpha}'/2 & \alpha_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}_{t-1} \\ 1 \end{bmatrix}. \tag{9}$$

The conditional variance h_t is positive if and only if the matrix in the quadratic form on the right-hand side of (9) is positive definite.

Some authors have suggested modelling the conditional standard deviation instead of the conditional variance: see Taylor (1986), Schwert (1990), and for an asymmetric version, Zakoïan (1994). A further generalization of this idea appeared in Ding et al. (1993). These authors proposed a GARCH model for h_t^k where $k > 0$ is also a parameter to be estimated. Their GARCH model is (2) with

$$h_t^k = \alpha_0 + \sum_{j=1}^{q} \alpha_j |\varepsilon_{t-j}|^{2k} + \sum_{j=1}^{p} \beta_j h_{t-j}^k, \ k > 0. \tag{10}$$

The authors argued that this extension provides flexibility lacking in the original GARCH specification of Bollerslev (1986) and Taylor (1986).

The proliferation of GARCH models has inspired some authors to define families of GARCH models that would accommodate as many individual

models as possible. Hentschel (1995) defined one such family . The first-order GARCH model has the general form

$$\frac{h_t^{\lambda/2}-1}{\lambda} = \omega + \alpha h_{t-1}^{\lambda/2} f^{\nu}(z_{t-1}) + \beta \frac{h_{t-1}^{\lambda/2}-1}{\lambda} \tag{11}$$

where $\lambda > 0$ and

$$f^{\nu}(z_t) = |z_t - b| - c(z_t - b).$$

Family (11) contains a large number of well-known GARCH models. The Box-Cox type transformation of the conditional standard deviation $h_t^{1/2}$ makes it possible, by allowing $\lambda \to 0$, to accommodate models in which the logarithm of the conditional variance is parameterized, such as the exponential GARCH model to be considered in Section 4. Parameters b and c in $f^{v}(z_t)$ allow the inclusion of different asymmetric GARCH models such as the GJR-GARCH or threshold GARCH models in (11).

Another family of GARCH models that is of interest is the one He and Teräsvirta (1999) defined as follows:

$$h_t^k = \sum_{j=1}^{q} g(z_{t-j}) + \sum_{j=1}^{p} c_j(z_{t-j}) h_{t-j}^k, \qquad k > 0 \tag{12}$$

where $\{g(z_t)\}$ and $\{c(z_t)\}$ are sequences of independent and identically distributed random variables. In fact, the family was originally defined for $q = 1$, but the definition can be generalized to higher-order models. For example, the standard GARCH(p, q) model is obtained by setting $g(z_t) = \alpha_0/q$ and $c_j(z_{t-j}) = \alpha_j z_{t-j}^2 + \beta_j, j = 1, \ldots, q$, in (12). Many other GARCH models such as the GJR-GARCH, the absolute-value GARCH, the Quadratic GARCH and the power GARCH model belong to this family.

Note that the power GARCH model itself nests several well-known GARCH models; see Ding et al. (1993) for details. Definition (12) has been used for deriving expressions of fourth moments, kurtosis and the autocorrelation function of ε_t^2 for a number of first-order GARCH models and the standard GARCH(p, q) model.

The family of augmented GARCH models, defined by Duan (1997), is a rather general family. The first-order augmented GARCH model is defined as follows. Consider (2) and assume that

$$h_t = \begin{cases} |\lambda\phi_t - \lambda - 1| \text{ if } \lambda \neq 0 \\ \exp\{\phi_t - 1\} \text{ if } \lambda = 0 \end{cases} \tag{13}$$

where

$$\phi_t = \alpha_0 + \zeta_{1,t-1}\phi_{t-1} + \zeta_{2,t-1}. \tag{14}$$

In (14), (ζ_{1t}, ζ_{2t}) is a strictly stationary sequence of random vectors with a continuous distribution, measurable with respect to the available information

until t. Duan defined an augmented GARCH(1,1) process as (2) with (13) and (14), such that

$$\zeta_{1t} = \alpha_1 + \alpha_2|\varepsilon_t - c|^\delta + \alpha_3 \max(0, c - \varepsilon_t)^\delta$$
$$\zeta_{2t} = \alpha_4 \frac{|\varepsilon_t - c|^\delta - 1}{\delta} + \alpha_5 \frac{\max(0, c - \varepsilon_t)^\delta - 1}{\delta}.$$

This process contains as special cases all the GARCH models previously mentioned, as well as the Exponential GARCH model to be considered in Section 4. Duan (1997) generalized this family to the GARCH(p, q) case and derived sufficient conditions for strict stationarity for this general family as well as conditions for the existence of the unconditional variance of ε_t. Furthermore, he suggested misspecification tests for the augmented GARCH model.

3.3 Nonlinear GARCH

3.3.1 Smooth transition GARCH

As mentioned above, the GARCH model has been extended to characterize asymmetric responses to shocks. The GJR-GARCH model, obtained as setting $\sum_{j=1}^{q} g(z_{t-j}) = \alpha_0$ and $c_j(z_{t-j}) = (\alpha_j + \omega_j I(z_{t-j} > 0))z_{t-j}^2 + \beta_j, j = 1, \ldots, q$, in (12), is an example of that. A nonlinear version of the GJR-GARCH model is obtained by making the transition between regimes smooth. Hagerud (1997), Gonzales-Rivera (1998) and Anderson et al. (1999) proposed this extension. A smooth transition GARCH (STGARCH) model may be defined as equation (2) with

$$h_t = \alpha_{10} + \sum_{j=1}^{q} \alpha_{1j}\varepsilon_{t-j}^2 + (\alpha_{20} + \sum_{j=1}^{q} \alpha_{2j}\varepsilon_{t-j}^2)G(\gamma, c; \varepsilon_{t-j}) + \sum_{j=1}^{p} \beta_j h_{t-j} \quad (15)$$

where the transition function

$$G(\gamma, c; \varepsilon_{t-j}) = (1 + \exp\{-\gamma \prod_{k=1}^{K} (\varepsilon_{t-j} - c_k)\})^{-1}, \ \gamma > 0. \quad (16)$$

When $K = 1$, (16) is a simple logistic function that controls the change of the coefficient of ε_{t-j}^2 from α_{1j} to $\alpha_{1j} + \alpha_{2j}$ as a function of ε_{t-j}, and similarly for the intercept. In that case, as $\gamma \to \infty$, the transition function becomes a step function and represents an abrupt switch from one regime to the other. Furthermore, at the same time setting $c_1 = 0$ yields the GJR-GARCH model because ε_t and z_t have the same sign. When $K = 2$ and, in addition, $c_1 = -c_2$ in (16), the resulting smooth transition GARCH model is still symmetric about zero, but the response of the conditional variance to a

shock is a nonlinear function of lags of ε_t^2. Smooth transition GARCH models are useful in situations where the assumption of two distinct regimes is not an adequate approximation to the asymmetric behaviour of the conditional variance. Hagerud (1997) also discussed a specification strategy that allows the investigator to choose between $K = 1$ and $K = 2$ in (16). Values of $K > 2$ may also be considered, but they are likely to be less common in applications than the two simplest cases.

The smooth transition GARCH model (15) with $K = 1$ in (16) is designed for modelling asymmetric responses to shocks. On the other hand, the standard GARCH model has the undesirable property that the estimated model often exaggerates the persistence in volatility (the estimated sum of the α- and β-coefficients is close to one). This in turn results in poor volatility forecasts. In order to remedy this problem, Lanne and Saikonnen (2005) proposed a smooth transition GARCH model whose first-order version has the form

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \delta_1 G_1(\boldsymbol{\theta}; h_{t-1}) + \beta_1 h_{t-1}. \tag{17}$$

In (17), $G_1(\boldsymbol{\theta}; h_{t-1})$ is a continuous bounded function such as (16): Lanne and Saikkonen use the cumulative distribution function of the gamma-distribution. A major difference between (15) and (17) is that in the latter model the transition variable is a lagged conditional variance. In empirical examples given in the paper, this parameterization clearly alleviates the problem of exaggerated persistence. The model may also be generalized to include a term of the form $G_1(\boldsymbol{\theta}; h_{t-1})h_{t-1}$, but according to the authors, such an extension appeared unnecessary in practice.

3.3.2 Threshold GARCH and extensions

If (15) is defined as a model for the conditional standard deviation such that h_t is replaced by $h_t^{1/2}$, h_{t-j} by $h_{t-j}^{1/2}$, $j = 1, \ldots, p$, and ε_{t-j}^2 by $|\varepsilon_{t-j}|$, $j = 1, \ldots, q$, then choosing $K = 1$, $c_1 = 0$ and letting $\gamma \to \infty$ in (16) yields the threshold GARCH (TGARCH) model that Zakoïan (1994) considered. The TGARCH(p, q) model is thus the counterpart of the GJR-GARCH model in the case where the entity to be modelled is the conditional standard deviation instead of the conditional variance. Note that in both of these models, the threshold parameter has a known value (zero). In Zakoïan (1994), the conditional standard deviation is defined as follows:

$$h_t^{1/2} = \alpha_0 + \sum_{j=1}^{q} (\alpha_j^+ \varepsilon_{t-j}^+ - \alpha_j^- \varepsilon_{t-j}^-) + \sum_{j=1}^{q} \beta_j h_{t-j}^{1/2} \tag{18}$$

where $\varepsilon_{t-j}^+ = \max(\varepsilon_{t-j}, 0)$, $\varepsilon_{t-j}^- = \min(\varepsilon_{t-j}, 0)$, and α_j^+, α_j^-, $j = 1, \ldots, q$, are parameters. Rabemananjara and Zakoïan (1993) introduced an even more general model in which $h_t^{1/2}$ can obtain negative values, but it has not gained

wide acceptance. Nevertheless, these authors provide evidence of asymmetry in the French stock market by fitting the TGARCH model (18) to the daily return series of stocks included in the CAC 40 index of the Paris Bourse.

The TGARCH model is linear in parameters because the threshold parameter is assumed to equal zero. A genuine nonlinear threshold model is the Double Threshold ARCH (DTARCH) model of Li and Li (1996). It is called a double threshold model because both the autoregressive conditional mean and the conditional variance have a threshold-type structure as defined in Tong (1990). The conditional mean is defined as follows:

$$y_t = \sum_{k=1}^{K}(\phi_{0k} + \sum_{j=1}^{p_k}\phi_{jk}y_{t-j})I(c_{k-1}^{(\mathrm{m})} < y_{t-b} \leq c_k^{(\mathrm{m})}) + \varepsilon_t \tag{19}$$

and the conditional variance has the form

$$h_t = \sum_{\ell=1}^{L}(\alpha_{0\ell} + \sum_{j=1}^{p_\ell}\alpha_{j\ell}\varepsilon_{t-j}^2)I(c_{\ell-1}^{(\mathrm{v})} < y_{t-d} \leq c_\ell^{(\mathrm{v})}). \tag{20}$$

Furthermore, $k = 1, \ldots, K$, $\ell = 1, \ldots, L$, and b and d are delay parameters, $b, d \geq 1$. The number of regimes in (19) and (20), K and L, respectively, need not be the same, nor do the two threshold variables have to be equal. Other threshold variables than lags of y_t are possible. For example, replacing y_{t-d} in (20) by ε_{t-d} or ε_{t-d}^2 may sometimes be an interesting possibility.

Another variant of the DTARCH model is the model that Audrino and Bühlmann (2001) who introduced it called the Tree-Structured GARCH model. It has an autoregressive conditional mean:

$$y_t = \phi y_{t-1} + \varepsilon_t \tag{21}$$

where ε_t is decomposed as in (2), and the first-order conditional variance

$$h_t = \sum_{k=1}^{K}(\alpha_{0k} + \alpha_{1k}y_{t-1}^2 + \beta_k h_{t-1})I\{(y_{t-1}, h_{t-1}) \in \mathcal{R}_k\}. \tag{22}$$

In (22), $\mathcal{R}_k$ is a subset in a partition $\mathcal{P} = \{\mathcal{R}_1, \ldots, \mathcal{R}_K\}$ of the sample space of (y_{t-1}, h_{t-1}). For example, if $K = 2$, either $\mathcal{R}_1 = \{y_{t-1} > c_y, h_{t-1} > 0\}$ or $\mathcal{R}_1 = \{-\infty < y_{t-1} < \infty, h_{t-1} > c_h\}$, $c_h > 0$, and $\mathcal{R}_2$ is the complement of $\mathcal{R}_1$. Note that, strictly speaking, equation (22) does not define a GARCH model unless $\phi = 0$ in (21), because the squared variable in the equation is y_{t-1}^2, not ε_{t-1}^2. A practical problem is that the tree-growing strategy of Audrino and Bühlmann (2001) does not seem to prevent underidentification: if K is chosen too large, (22) is not identified. A similar problem is present in the DTARCH model as well as in the STGARCH one. Hagerud (1997) and Gonzales-Rivera (1998), however, do provide linearity tests in order to avoid this problem in the STGARCH framework.

3.4 Time-varying GARCH

An argument brought forward in the literature, see for instance Mikosch and Stărică (2004), is that in applications the assumption of the GARCH models having constant parameters may not be appropriate when the series to be modelled are long. Parameter constancy is a testable proposition, and if it is rejected, the model can be generalized. One possibility is to assume that the parameters change at specific points of time, divide the series into subseries according to the location of the break-points, and fit separate GARCH models to the subseries. The main statistical problem is then finding the number of break-points and their location because they are normally not known in advance. Chu (1995) has developed tests for this purpose.

Another possibility is to modify the smooth transition GARCH model (15) to fit this situation. This is done by defining the transition function (16) as a function of time:

$$G(\gamma, c; t^*) = (1 + \exp\{-\gamma \prod_{k=1}^{K} (t^* - c_k)\})^{-1}, \ \gamma > 0$$

where $t^* = t/T$, $t = 1, \ldots, T$, and T is the number of observations. Standardizing the time variable between zero and unity makes interpretation of the parameters $c_k, k = 1, \ldots, K$, easy as they indicate where in relative terms the changes in the process occur. The resulting time-varying parameter GARCH (TV-GARCH) model has the form

$$h_t = \alpha_0(t) + \sum_{j=1}^{q} \alpha_j(t)\varepsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j(t) h_{t-j} \tag{23}$$

where $\alpha_0(t) = \alpha_{01} + \alpha_{02}G(\gamma, c; t^*)$, $\alpha_j(t) = \alpha_{j1} + \alpha_{j2}G(\gamma, c; t^*), j = 1, \ldots, q$, and $\beta_j(t) = \beta_{j1} + \beta_{j2}G(\gamma, c; t^*)$, $j = 1, \ldots, p$. This is the most flexible parameterization. Some of the time-varying parameters in (23) may be restricted to constants *a priori*. For example, it may be assumed that only the intercept $\alpha_0(t)$ is time-varying. This implies that the 'baseline volatility' or unconditional variance is changing over time. If change is allowed in the other GARCH parameters then the model is capable of accommodating systematic changes in the amplitude of the volatility clusters that cannot be explained by a constant-parameter GARCH model.

This type of time-varying GARCH is mentioned here because it is a special case of the smooth transition GARCH model. Other time-varying parameter models of conditional heteroskedasticity, such as nonstationary ARCH models with locally changing parameters, are discussed in Čížek and Spokoiny (2008).

3.5 Markov-switching ARCH and GARCH

Markov-switching or hidden Markov models of conditional heteroskedasticity constitute another class of nonlinear models of volatility. These models are an alternative way of modelling volatility processes that contains breaks. Hamilton and Susmel (1994) argued that very large shocks, such as the one affecting the stocks in October 1987, may have consequences for subsequent volatility so different from consequences of small shocks that a standard ARCH or GARCH model cannot describe them properly. Their Markov-switching ARCH model is defined as follows:

$$h_t = \sum_{i=1}^{k} (\alpha_0^{(i)} + \sum_{j=1}^{q} \alpha_j^{(i)} \varepsilon_{t-j}^2) I(s_t = i) \tag{24}$$

where s_t is a discrete unobservable random variable obtaining values from the set $S = \{1, \ldots, k\}$ of regime indicators. It follows a (usually first-order) homogeneous Markov chain:

$$\Pr\{s_t = j | s_t = i\} = p_{ij}, \ i, j = 1, \ldots, k.$$

Cai (1994) considered a special case of (24) in which only the intercept $\alpha_0^{(i)}$ is switching, and $k = 2$. But then, his model also contains a switching conditional mean. Furthermore, Rydén et al. (1998) showed that a simplified version of (24) where $\alpha_j^{(i)} = 0$ for $j \geq 1$ and all i, is already capable of generating data that display most of the stylized facts that Granger and Ding (1995) ascribe to high-frequency, daily, say, financial return series. This suggests that a Markov-switching variance alone without any ARCH structure may in many cases explain a large portion of the variation in these series.

Nevertheless, it can be argued that shocks drive economic processes, and this motivates the ARCH structure. If the shocks have a persistent effect on volatility, however, a parsimonious GARCH representation may be preferred to (24). Generalizing (24) into a GARCH model involves one major difficulty. A straightforward (first-order) generalization would have the following form:

$$h_t = (\alpha_0^{(i)} + \alpha_1^{(i)} \varepsilon_{t-1}^2 + \beta_1^{(i)} h_{t-1}) I(s_t = i). \tag{25}$$

From the autoregressive structure of (25) it follows that h_t is completely path-dependent: its value depends on the unobservable s_{t-j}, $j = 0, 1, 2, \ldots t$. This makes the model practically impossible to estimate because in order to evaluate the log-likelihood, these unobservables have to be integrated out of this function. Simplifications of the model that circumvent this problem can be found in Gray (1996) and Klaassen (2002). A good discussion about their models can be found in Haas et al. (2004). These authors present another MS-GARCH model whose fourth-moment structure they are able to work out. That does not seem possible for the other models. The MS-GARCH

model of Haas et al. (2004) is defined as follows:

$$\varepsilon_t = z_t \sum_{i=1}^{k} h_{it}^{1/2} I(s_t = i)$$

where s_t is defined as in (24). Furthermore,

$$\boldsymbol{h}_t = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 \varepsilon_{t-1}^2 + \boldsymbol{B}\boldsymbol{h}_{t-1}$$

where $\boldsymbol{h}_t = (h_{1t}, \ldots, h_{1k})'$, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ik})'$, $i = 0, 1$, and $\boldsymbol{B} = \mathrm{diag}(\beta_{11}, \ldots, \beta_{1k})'$. Thus, each volatility regime has its own GARCH equation. The conditional variance in a given regime is only a function of the lagged conditional variance in the same regime, which is not the case in the other models. The identification problem mentioned in Section 3.3.2 is present here as well. If the true model has fewer regimes than the specified one, the latter contains unidentified nuisance parameters.

More information about Markov-switching ARCH and GARCH models can be found in Lange and Rahbek (2008).

3.6 Integrated and fractionally integrated GARCH

In applications it often occurs that the estimated sum of the parameters α_1 and β_1 in the standard first-order GARCH model (4) with $p = q = 1$ is close to unity. Engle and Bollerslev (1986), who first paid attention to this phenomenon, suggested imposing the restriction $\alpha_1 + \beta_1 = 1$ and called the ensuing model an integrated GARCH (IGARCH) model. The IGARCH process is not weakly stationary as $\mathsf{E}\varepsilon_t^2$ is not finite. Nevertheless, the term "integrated GARCH" may be somewhat misleading as the IGARCH process is strongly stationary. Nelson (1991) showed that under mild conditions for $\{z_t\}$ and assuming $\alpha_0 > 0$, the GARCH(1,1) process is strongly stationary if

$$\mathsf{E}\ln(\alpha_1 + \beta_1 z_t^2) < 0 \tag{26}$$

(recall that $\mathsf{E}z_t^2 = 1$). The IGARCH process satisfies (26). The analogy with integrated processes, that is, ones with a unit root, is therefore not as straightforward as one might think. For a general discussion of stationarity conditions in GARCH models, see Lindner (2008).

Nelson (1991) also showed that when an IGARCH process is started at some finite time point, its behaviour depends on the intercept α_0. On the one hand, if the intercept is positive then the unconditional variance of the process grows linearly with time. In practice this means that the amplitude of the clusters of volatility to be parameterized by the model on the average increases over time. The rate of increase need not, however, be particularly

rapid. One may thus think that in applications with, say, a few thousand observations, IGARCH processes nevertheless provide a reasonable approximation to the true data-generating volatility process. On the other hand, if $\alpha_0 = 0$ in the IGARCH model, the realizations from the process collapse to zero almost surely. How rapidly this happens, depends on the parameter β_1.

Although the investigator may be prepared to accept an IGARCH model as an approximation, a potentially disturbing fact is that this means assuming that the unconditional variance of the process to be modelled does not exist. It is not clear that this is what one always wants to do. There exist other explanations to the fact that the sum $\alpha_1 + \beta_1$ estimates to one or very close to one. First Diebold (1986) and later Lamoureux and Lastrapes (1990) suggested that this often happens if there is a switch in the intercept of a GARCH model during the estimation period. This may not be surprising as such a switch means that the underlying GARCH process is not stationary.

Another, perhaps more puzzling, observation is related to exponential GARCH models to be considered in Section 4. Malmsten (2004) noticed that if a GARCH$(1,1)$ model is fitted to a time series generated by a stationary first-order exponential GARCH model (see Section 4), the probability of the estimated sum $\alpha_1 + \beta_1$ exceeding unity can sometimes be rather large. In short, if the estimated sum of these two parameters in a standard GARCH(1,1) model is close to unity, imposing the restriction $\alpha_1 + \beta_1 = 1$ without further investigation may not necessarily be the most reasonable action to take.

Assuming $p = q = 1$, the GARCH(p, q) equation (4) can also be written in the "ARMA(1,1) form" by adding ε_t^2 to both sides and moving h_t to the right-hand side:

$$\varepsilon_t^2 = \alpha_0 + (\alpha_1 + \beta_1)\varepsilon_{t-1}^2 + \nu_t - \beta_1 \nu_{t-1} \tag{27}$$

where $\{\nu_t\} = \{\varepsilon_t^2 - h_t\}$ is a martingale difference sequence with respect to h_t. For the IGARCH process, (27) has the "ARIMA(0,1,1) form"

$$(1 - L)\varepsilon_t^2 = \alpha_0 + \nu_t - \beta_1 \nu_{t-1}. \tag{28}$$

Equation (28) has served as a starting-point for the fractionally integrated GARCH (FIGARCH) model. The FIGARCH(1,d,0) model is obtained from (28) by replacing the difference operator by a fractional difference operator:

$$(1 - L)^d \varepsilon_t^2 = \alpha_0 + \nu_t - \beta_1 \nu_{t-1}. \tag{29}$$

The FIGARCH equation (29) can be written as an infinite-order ARCH model by applying the definition $\nu_t = \varepsilon_t^2 - h_t$ to it. This yields

$$h_t = \alpha_0 (1 - \beta_1)^{-1} + \lambda(L)\varepsilon_t^2$$

where $\lambda(L) = \{1 - (1-L)^d(1-\beta_1 L)^{-1}\}\varepsilon_t^2 = \sum_{j=1}^{\infty} \lambda_j L^j \varepsilon_t^2$, and $\lambda_j \geq 0$ for all j. Expanding the fractional difference operator into an infinite sum yields the result that for long lags j,

$$\lambda_j = \{(1-\beta_1)\Gamma(d)^{-1}\}j^{-(1-d)} = cj^{-(1-d)}, c > 0 \tag{30}$$

where $d \in (0,1)$ and $\Gamma(d)$ is the gamma function. From (30) it is seen that the effect of the lagged ε_t^2 on the conditional variance decays hyperbolically as a function of the lag length. This is the reason why Ballie et al. (1996) introduced the FIGARCH model, as it would conveniently explain the apparent slow decay in autocorrelation functions of squared observations of many daily return series. The FIGARCH model thus offers a competing view to the one according to which changes in parameters in a GARCH model are the main cause of the slow decay in the autocorrelations. The first-order FIGARCH model (29) can of course be generalized into a FIGARCH(p,d,q) model.

The probabilistic properties of FIGARCH processes such as stationarity, still an open question, are quite complex, see, for example, Davidson (2004) and Giraitis et al. (2008) for discussion. The hyperbolic GARCH model introduced in the first-mentioned paper contains the standard GARCH and the FIGARCH models as two extreme special cases; for details see Davidson (2004).

3.7 Semi- and nonparametric ARCH models

The ARCH decomposition of returns (2) has also been used in a semi- or nonparametric approach. The semiparametric approach is typically employed in situations where the distribution of z_t is left unspecified and is estimated nonparametrically. In nonparametric models, the issue is the estimation of the functional form of the relationship between ε_t^2 and $\varepsilon_{t-1}^2, \ldots, \varepsilon_{t-q}^2$. Semi- and nonparametric ARCH models are considered in detail in Linton (2008).

3.8 GARCH-in-mean model

GARCH models are often used for predicting the risk of a portfolio at a given point of time. From this it follows that the GARCH type conditional variance could be useful as a representation of the time-varying risk premium in explaining excess returns, that is, returns compared to the return of a riskless asset. An excess return would be a combination of the unforecastable difference ε_t between the ex ante and ex post rates of return and a function of the conditional variance of the portfolio. Thus, if y_t is the excess return at time t,

$$y_t = \varepsilon_t + \beta + g(h_t) - \mathsf{E}g(h_t) \tag{31}$$

where h_t is defined as a GARCH process (4) and $g(h_t)$ is a positive-valued function. Engle et al. (1987) originally defined $g(h_t) = \delta h_t^{1/2}, \delta > 0$, which corresponds to the assumption that changes in the conditional standard deviation appear less than proportionally in the mean. The alternative $g(h_t) = \delta h_t$ has also appeared in the literature. Equations (31) and (4) form the GARCH-in-mean or GARCH-M model. It has been quite frequently applied in the applied econometrics and finance literature. Glosten et al. (1993) developed their asymmetric GARCH model as a generalization of the GARCH-M model.

The GARCH-M process has an interesting moment structure. Assume that $\mathsf{E}z_t^3 = 0$ and $\mathsf{E}\varepsilon_t^4 < \infty$. From (31) it follows that the kth order autocovariance

$$\mathsf{E}(y_t - \mathsf{E}y_t)(y_{t-k} - \mathsf{E}y_t) = \mathsf{E}\varepsilon_{t-k}g(h_t) + \operatorname{cov}(g(h_t), g(h_{t-k})) \neq 0.$$

This means that there is forecastable structure in y_t, which may contradict some economic theory if y_t is a return series. Hong (1991) showed this in a special case where $g(h_t) = \delta h_t$, $\mathsf{E}\varepsilon_t^4 < \infty$, and h_t follows a GARCH(p,q) model. In that situation, all autocorrelations of y_t are nonzero. Furthermore,

$$\mathsf{E}(y_t - \mathsf{E}y_t)^3 = 3\mathsf{E}h_t\{g(h_t) - \mathsf{E}g(h_t)\} + \mathsf{E}\{g(h_t) - \mathsf{E}g(h_t)\}^3 \neq 0. \tag{32}$$

It follows from (32) that a GARCH-M model implies postulating a skewed marginal distribution for y_t unless $g(h_t) \equiv$ constant. For example, if $g(h_t) = \delta h_t^{1/2}$, $\delta < 0$, this marginal distribution is negatively skewed. If the model builder is not prepared to make this assumption or the one of forecastable structure in y_t, the GARCH-M model, despite its theoretical motivation, does not seem an appropriate alternative to use. For more discussion of this situation, see He et al. (2006).

3.9 Stylized facts and the first-order GARCH model

As already mentioned, financial time series such as high-frequency return series constitute the most common field of applications for GARCH models. These series typically display rather high kurtosis. At the same time, the autocorrelations of the absolute values or squares of the observations are low and decay slowly. These two features are sometimes called stylized facts of financial time series. Granger and Ding (1995) listed a few more such features. Among them is the empirical observation that in a remarkable number of financial series, the autocorrelations of the powers of observations, $|\varepsilon_t|^k$, peak around $k = 1$. Granger and Ding called this stylized fact the Taylor effect as Taylor (1986) was the first to draw attention to it (by comparing the autocorrelations of ε_t^2 and $|\varepsilon_t|$).

One way of evaluating the adequacy of GARCH models is to ask how well they can be expected to capture the features or stylized facts present in the series to be modelled. The expressions for kurtosis and the autocorrelation function of absolute values and squared observations are available for the purpose. They allow one to find out, for example, whether or not a GARCH(1,1) model is capable of producing realizations with high kurtosis and low, slowly decaying autocorrelations. The results of Malmsten and Teräsvirta (2004) who have used these expressions, illustrate the well known fact, see, for example, Bollerslev et al. (1994), that a GARCH model with normally distributed errors does not seem to be a sufficiently flexible model for explaining these two features in financial return series. This is shown in Figure 1. The panels contain a number of isoquants for which the sum

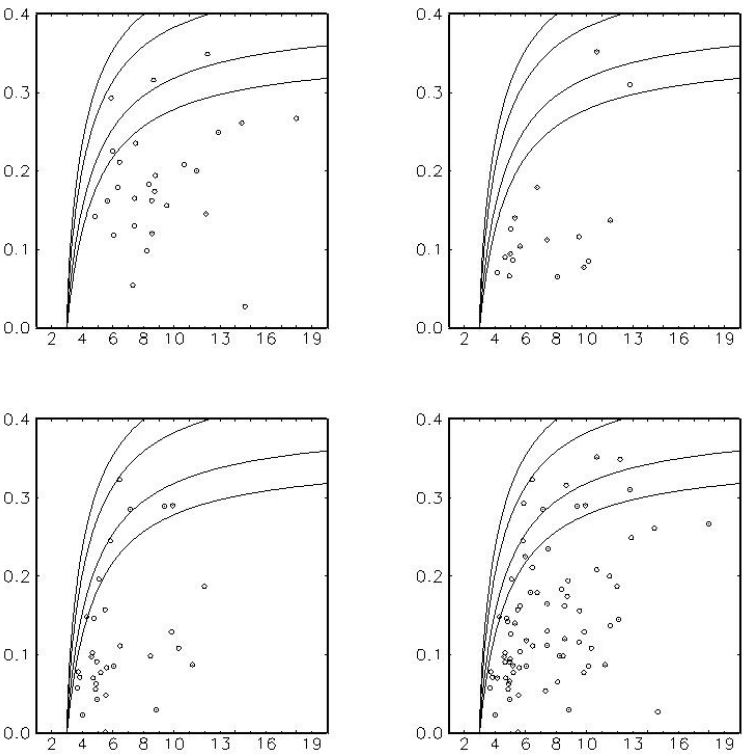

Fig. 1 Kurtosis/first-order autocorrelation isoquants for the GARCH(1,1) model, from highest to lowest: $\alpha + \beta = 0.9, 0.95, 0.99, 0.999$, and corresponding combinations estimated from data: Upper left panel: Daily returns of the 27 most actively traded stocks at the Stockholm Stock Exchange; Upper right panel: Returns of five major daily exchange rates, divided into 34 subseries; Lower left panel: Daily returns of the S&P 500 index from 3 January 1928 to 19 September 2001, divided into 20 equally long subseries; Lower right panel: All observations

$\alpha_1 + \beta_1$ remains constant as a function of the kurtosis and the first-order autocorrelation of squared observations. Note that $\alpha_1 + \beta_1$ is the exponential decay rate of the autocorrelation function, that is, the jth autocorrelation $\rho_j = (\alpha_1 + \beta_1)^{j-1}\rho_1$ for $j \geq 1$. They also contain combinations of the kurtosis and the first-order autocorrelation estimated directly from time series. It is seen that very often the kurtosis/autocorrelation combinations do not tend to lie in the vicinity of these isoquants even when $\alpha_1 + \beta_1$ is very close to one. The isoquants are od course only defined for combinations of α_1 and β_1 for which $E\varepsilon_t^4 < \infty$.

Malmsten and Teräsvirta (2004) also demonstrated how the situation can be improved, as is customary in practice, by replacing the normal error distribution by a more fat-tailed one. In Figure 2 it is seen how increasing the "baseline kurtosis", that is, the kurtosis of the distribution of z_t, the error, helps the GARCH(1,1) model to capture the stylized fact of high kurtosis/low autocorrelation. The isoquants are moved to the right because the baseline

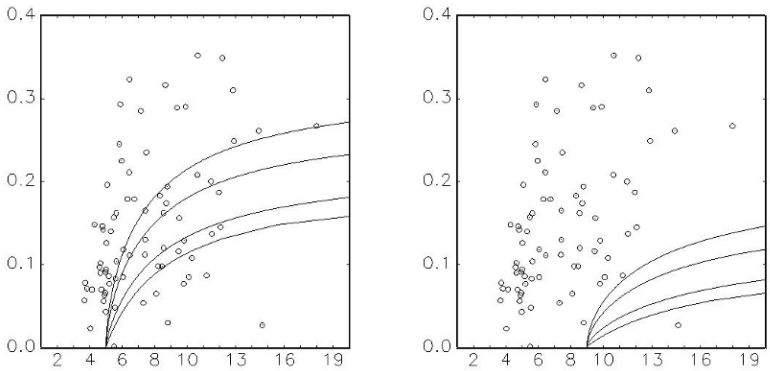

Fig. 2 Isoquants of pairs of kurtosis and first-order autocorrelation of squared observations in the GARCH(1,1) model with $t(7)$-distributed (left-hand panel) and $t(5)$-distributed errors (right-hand panel), for (from above) $\alpha + \beta = 0.90, 0.95, 0.99$ and 0.999, and corresponding observations (the same ones as in the lower right panel of Figure 1).

kurtosis increases. At the same time it seems that the level of the autocorrelations decreases. But then, this does not simultaneously affect the decay rate $\alpha_1 + \beta_1$ of the autocorrelations.

Recently, Kim and White (2004) suggested that the standard estimator of kurtosis exaggerates the true kurtosis and that robust measures yield more reliable results. It follows that high kurtosis values estimated from return series are a result of a limited number of outliers. If this is the case, then the use of a non-normal (heavy-tailed) error distribution may not necessarily be an optimal extension to the standard normal-error GARCH model. However, Teräsvirta and Zhao (2006) recently studied 160 daily return series and, following Kim and White (2004), used robust kurtosis and autocorrelation

estimates instead of standard ones. Their results indicate that leptokurtic distributions for z_t are needed in capturing the kurtosis-autocorrelation stylized fact even when the influence of extreme observations is dampened by the use of robust estimates.

As to the Taylor effect, He and Teräsvirta (1999) defined a corresponding theoretical property, the Taylor property, as follows. Let $\rho(|\varepsilon_t|^k, |\varepsilon_{t-j}|^k)$ be the jth order autocorrelation of $\{|\varepsilon_t|^k\}$. The stochastic process has the Taylor property when $\rho(|\varepsilon_t|^k, |\varepsilon_{t-j}|^k)$ is maximized for $k = 1$ for $j = 1, 2, \ldots$. In practice, He and Teräsvirta (1999) were able to find analytical results for the AVGARCH(1,1) model, but they were restricted to comparing the first-order autocorrelations for $k = 1$ and $k = 2$. For this model, $\rho(|\varepsilon_t|, |\varepsilon_{t-1}|) > \rho(\varepsilon_t^2, \varepsilon_{t-1}^2)$ when the kurtosis of the process is sufficiently high. The corresponding results for the standard GARCH(1,1) model (4) with $p = q = 1$ and normal errors are not available as the autocorrelation function of $\{|\varepsilon_t|\}$ cannot be derived analytically. Simulations conducted by He and Teräsvirta (1999) showed that the GARCH(1,1) model probably does not possess the Taylor property, which may seem disappointing. But then, the results of Teräsvirta and Zhao (2006) show that if the standard kurtosis and autocorrelation estimates are replaced by robust ones, the evidence of the Taylor effect completely disappears. This stylized fact may thus be a consequence of just a small number of extreme observations in the series.

4 Family of Exponential GARCH Models

4.1 Definition and properties

The Exponential GARCH (EGARCH) model is another popular GARCH model. Nelson (1991) who introduced it had three criticisms of the standard GARCH model in mind. First, parameter restrictions are required to ensure positivity of the conditional variance at every point of time. Second, the standard GARCH model does not allow an asymmetric response to shocks. Third, if the model is an IGARCH one, measuring the persistence is difficult since this model is strongly but not weakly stationary. Shocks may be viewed persistent as the IGARCH process looks like a random walk. However, the IGARCH model with $\alpha_0 > 0$ is strictly stationary and ergodic, and when $\alpha_0 = 0$, the realizations collapse into zero almost surely, as already indicated in Section 3.6. The second drawback has since been removed as asymmetric GARCH models such as GJR-GARCH (Glosten et al. (1993)) or smooth transition GARCH have become available. A family of EGARCH(p, q) models may be defined as in (2) with

$$\ln h_t = \alpha_0 + \sum_{j=1}^{q} g_j(z_{t-j}) + \sum_{j=1}^{p} \beta_j \ln h_{t-j}. \tag{33}$$

When $g_j(z_{t-j}) = \alpha_j z_{t-j} + \psi_j(|z_{t-j}| - \mathsf{E}|z_{t-j}|), j = 1, \ldots, q$, (33) becomes the EGARCH model of Nelson (1991). It is seen from (33) that no parameter restrictions are necessary to ensure positivity of h_t. Parameters $\alpha_j, j = 1, \ldots, q$, make an asymmetric response to shocks possible.

When $g_j(z_{t-j}) = \alpha_j \ln z_{t-j}^2$, $j = 1, \ldots, q$, (2) and (33) form the logarithmic GARCH (LGARCH) model that Geweke (1986) and Pantula (1986) proposed. The LGARCH model has not become popular among practitioners. A principal reason for this may be that for parameter values encountered in practice, the theoretical values of the first few autocorrelations of $\{\varepsilon_t^2\}$ at short lags tend to be so high that such autocorrelations can hardly be found in financial series such as return series. This being the case, the LGARCH model cannot be expected to provide an acceptable fit when applied to financial series. Another reason are the occasional small values of $\ln \varepsilon_t^2$ that complicate the estimation of parameters.

As in the standard GARCH case, the first-order model is the most popular EGARCH model in practice. Nelson (1991) derived existence conditions for moments of the general infinite-order Exponential ARCH model. Translated to the case of the EGARCH model (2) and (33) such that $g_j(z_{t-j}) = \alpha_j z_{t-j} + \psi_j(|z_{t-j}| - \mathsf{E}|z_{t-j}|)$, $j = 1, \ldots, q$, where not all α_j and ψ_j equal zero, these existence conditions imply that if the error process $\{z_t\}$ has all moments and $\sum_{j=1}^{p} \beta_j^2 < 1$ in (33), then all moments for the EGARCH process $\{\varepsilon_t\}$ exist. For example, if $\{z_t\}$ is a sequence of independent standard normal variables then the restrictions on β_j, $j = 1, \ldots, p$, are necessary and sufficient for the existence of all moments simultaneously. This is different from the family (12) of GARCH models considered in Section 3.2. For those models, the moment conditions become more and more stringent for higher and higher even moments. The expressions for moments of the first-order EGARCH process can be found in He et al. (2002); for the more general case, see He (2000).

4.2 Stylized facts and the first-order EGARCH model

In Section 3.9 we considered the capability of first-order GARCH models to characterize certain stylized facts in financial time series. It is instructive to do the same for EGARCH models. For the first-order EGARCH model, the decay of autocorrelations of squared observations is faster than exponential in the beginning before it slows down towards an exponential rate; see He et al. (2002). Thus it does not appear possible to use a standard EGARCH(1,1) model to characterize processes with very slowly decaying autocorrelations. Malmsten and Teräsvirta (2004) showed that the symmetric EGARCH(1,1)

model with normal errors is not sufficiently flexible either for characterizing series with high kurtosis and slowly decaying autocorrelations. As in the standard GARCH case, assuming normal errors means that the first-order autocorrelation of squared observations increases quite rapidly as a function of kurtosis for any fixed β_1 before the increase slows down. Analogously to GARCH, the observed kurtosis/autocorrelation combinations cannot be reached by the EGARCH(1,1) model with standard normal errors. The asymmetry parameter is unlikely to change things much.

Nelson (1991) recommended the use of the so-called Generalized Error Distribution (GED) for the errors. It contains the normal distribution as a special case but also allows heavier tails than the ones in the normal distribution. Nelson (1991) also pointed out that a t-distribution for the errors may imply infinite unconditional variance for $\{\varepsilon_t\}$. As in the case of the GARCH(1,1) model, an error distribution with fatter tails than the normal one helps to increase the kurtosis and, at the same time, lower the autocorrelations of squared observations or absolute values.

Because of analytical expressions of the autocorrelations for $k > 0$ given in He et al. (2002) it is possible to study the existence of the Taylor property in EGARCH models. Using the formulas for the autocorrelations of $\{|\varepsilon_t|^k\}$, $k > 0$, it is possible to find parameter combinations for which these autocorrelations peak in a neighbourhood of $k = 1$. A subset of first-order EGARCH models thus has the Taylor property. This subset is also a relevant one in practice in the sense that it contains EGARCH processes with the kurtosis of the magnitude frequently found in financial time series. For more discussion on stylized facts and the EGARCH(1,1) model, see Malmsten and Teräsvirta (2004).

4.3 Stochastic volatility

The EGARCH equation may be modified by replacing $g_j(z_{t-j})$ by $g_j(s_{t-j})$ where $\{s_t\}$ is a sequence of continuous unobservable independent random variables that are often assumed independent of z_t at all lags. Typically in applications, $p = q = 1$ and $g_1(s_{t-1}) = \delta s_{t-1}$ where δ is a parameter to be estimated. This generalization is called the autoregressive stochastic volatility (SV) model, and it substantially increases the flexibility of the EGARCH parameterization. For evidence of this, see Malmsten and Teräsvirta (2004) and Carnero et al. (2004). A disadvantage is that model evaluation becomes more complicated than that of EGARCH models because the estimation does not yield residuals. Several articles in this *Handbook* are devoted to SV models.

5 Comparing EGARCH with GARCH

The standard GARCH model is probably the most frequently applied parameterization of conditional heteroskedasticity. This being the case, it is natural to evaluate an estimated EGARCH model by testing it against the corresponding GARCH model. Since the EGARCH model can characterize asymmetric responses to shocks, a GARCH model with the same property, such as the GJR-GARCH or the smooth transition GARCH model, would be a natural counterpart in such a comparison. If the aim of the comparison is to choose between these models, they may be compared by an appropriate model selection criterion as in Shephard (1996). Since the GJR-GARCH and the EGARCH model of the same order have equally many parameters, this amounts to comparing their maximized likelihoods.

If the investigator has a preferred model or is just interested in knowing if there are significant differences in the fit between the two, the models may be tested against each other. The testing problem is a non-standard one because the two models do not nest each other. Several approaches have been suggested for this situation. Engle and Ng (1993) proposed combining the two models into an encompassing model. If the GARCH model is an GJR-GARCH(p, q) one (both models can account for asymmetries), this leads to the following specification of the conditional variance:

$$\ln h_t = \sum_{j=1}^{q} \{\alpha_j^* z_{t-j} + \psi_j^* (|z_{t-j}| - \mathsf{E}|z_{t-j}|)\} + \sum_{j=1}^{p} \beta_j^* \ln h_{t-j} + \ln(\alpha_0 + \sum_{j=1}^{q} \{\alpha_j + \omega_j I(\varepsilon_{t-j})\} \varepsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}). \tag{34}$$

Setting $(\alpha_j, \omega_j) = (0, 0), j = 1, \ldots, q$, and $\beta_j = 0$, $j = 1, \ldots, p$, in (34) yields an EGARCH(p, q) model. Correspondingly, the restrictions $(\alpha_j^*, \psi_j^*) = (0, 0)$, $j = 1, \ldots, q$, and $\beta_j^* = 0$, $j = 1, \ldots, p$, define the GJR-GARCH(p, q) model. Testing the models against each other amounts to testing the appropriate restrictions in (34). A Lagrange Multiplier test may be constructed for the purpose. The test may also be viewed as another misspecification test and not only as a test against the alternative model.

Another way of testing the EGARCH model against GARCH consists of forming the likelihood ratio statistic despite the fact that the null model is not nested in the alternative. This is discussed in Lee and Brorsen (1997) and Kim et al. (1998). Let $\mathcal{M}_0$ be the EGARCH model and $\mathcal{M}_1$ the GARCH one, and let the corresponding log-likelihoods be $L_T(\boldsymbol{\varepsilon}; \mathcal{M}_0, \boldsymbol{\theta}_0)$ and $L_T(\boldsymbol{\varepsilon}; \mathcal{M}_1, \boldsymbol{\theta}_1)$, respectively. The test statistic is

$$LR = 2\{L_T(\boldsymbol{\varepsilon}; \mathcal{M}_1, \widehat{\boldsymbol{\theta}}_1) - L_T(\boldsymbol{\varepsilon}; \mathcal{M}_0, \widetilde{\boldsymbol{\theta}}_0)\}. \tag{35}$$

The asymptotic null distribution of (35) is unknown but can be approximated by simulation. Assuming that the EGARCH model is the null model and that $\widetilde{\boldsymbol{\theta}}_0$ is the true parameter, one generates N realizations of T observations from this model and estimates both models and calculates the value of (35) using each realization. Ranking the N values gives an empirical distribution with which one compares the original value of (35). The true value of $\boldsymbol{\theta}_0$ is unknown, but the approximation error due to the use of $\widetilde{\boldsymbol{\theta}}_0$ as a replacement vanishes asymptotically as $T \to \infty$. If the value of (35) exceeds the $100(1-\alpha)\%$ quantile of the empirical distribution, the null model is rejected at significance level α. Note that the models under comparison need not have the same number of parameters, and the value of the statistic can also be negative. Reversing the roles of the models, one can test GARCH models against EGARCH ones.

Chen and Kuan (2002) proposed yet another method based on the pseudo-score, whose estimator under the null hypothesis and assuming the customary regularity conditions is asymptotically normally distributed. This result forms the basis for a χ^2-distributed test statistic; see Chen and Kuan (2002) for details.

Results of small-sample simulations in Malmsten (2004) indicate that the pseudo-score test tends to be oversized. Furthermore, the Monte Carlo likelihood ratio statistic seems to have consistently higher power than the encompassing test, which suggests that the former rather than the latter should be applied in practice.

6 Final Remarks and Further Reading

The literature on univariate GARCH models is quite voluminous, and it is not possible to incorporate all developments and extensions of the original model in the present text. Several articles of this *Handbook* provide detailed analyses of various aspects of GARCH models. Modern econometrics texts contain accounts of conditional heteroskedasticity. A number of surveys of GARCH models exist as well. Bollerslev et al. (1994), Diebold and Lopez (1995), Palm (1996), and Guégan (1994) (Chapter 5) survey developments till the early 1990s; see Giraitis et al. (2006) for a very recent survey. Shephard (1996) considers both univariate GARCH and stochastic volatility models. The focus in Gouriéroux (1996) lies on both univariate and multivariate ARCH models. The survey by Bollerslev et al. (1992) also reviews applications to financial series. The focus in Straumann (2004) is on estimation in models of conditional heteroskedasticity. Theoretical results on time series models with conditional heteroskedasticity are also reviewed in Li et al. (2002). Engle (1995) contains a selection of the most important articles on ARCH and GARCH models up until 1993.

Multivariate GARCH models are not included in this article. There exists a recent survey by Bauwens et al. (2006), and these models are also considered in Silvennoinen and Teräsvirta (2008).

Acknowledgement This research has been supported by Jan Wallander's and Tom Hedelius's Foundation, Grant No. P2005-0033:1. A part of the work for the chapter was done when the author was visiting Sonderforschungsbereich 649 at the Humboldt University Berlin. Comments from Changli He, Marcelo Medeiros and Thomas Mikosch (editor) are gratefully acknowledged. Any shortcomings and errors are the author's sole responsibility.

References

Andersen, T.G. and Benzoni, L. (2008): Realized volatility. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 554–575. Springer, New York.

Andersen, T.G. and Bollerslev, T. (1998): Answering the Skeptics: Yes, Standard Volatility Models Provide Accurate Forecasts. *International Economic Review* **39**, 885–905.

Anderson, H.M., Nam, K. and Vahid, F. (1999): Asymmetric nonlinear smooth transition GARCH models. *In: Rothmann, P. (Ed.): Nonlinear time series analysis of economic and financial data,* 191–207. Kluwer, Boston.

Audrino, F. and Bühlmann, P. (2001): Tree-Structured Generalized Autoregressive Conditional Heteroscedastic Models. *Journal of the Royal Statistical Society B* **63**, 727–744.

Ballie R.T., Bollerslev, T. and Mikkelsen, H.O. (1996): Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **74**, 3–30.

Bauwens, L., Laurent, S. and Rombouts, J.V.K. (2006): Multivariate GARCH Models: A Survey. *Journal of Applied Econometrics* **21**, 79–109.

Bollerslev, T. (1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T., Chou, R.Y. and Kroner, K.F. (1992): ARCH Modeling in Finance. A Review of the Theory and Empirical Evidence. *Journal of Econometrics* **52**, 5–59.

Bollerslev, T., Engle, R.F. and Nelson, D.B. (1994): ARCH Models. *In: Engle, R.F. and McFadden, D.L. (Eds.): Handbook of Econometrics, volume 4*, 2959–3038. North-Holland, Amsterdam.

Cai, J. (1994): A Markov Model of Switching-Regime ARCH. *Journal of Business and Economic Statistics* **12**, 309–316.

Carnero, M.A., Peña, D. and Ruiz, E. (2004): Persistence and Kurtosis in GARCH and Stochastic Volatility Models. *Journal of Financial Econometrics* **2**, 319–342.

Chen, Y.-T. and Kuan, C.-M. (2002): The Pseudo-True Score Encompassing Test for Non-Nested Hypotheses. *Journal of Econometrics* **106**, 271–295.

Chu, C.-S.J.C. (1995): Detecting Parameter Shift in GARCH Models. *Econometric Reviews* **14**, 241–266.

Čížek, P. and Spokoiny, V. (2008): Varying coefficient GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 168–185. Springer, New York.

Davidson, J. (2004): Moment and Memory Properties of Linear Conditional Heteroscedasticity Models, and a New Model. *Journal of Business and Economic Statistics* **22**, 16–29.

Diebold, F.X. and Lopez, J.A. (1995): Modeling Volatility Dynamics. *In: Hoover, K.D. (Ed.): Macroeconometrics: Developments, Tensions, and Prospects*, 427–472. Kluwer, Boston.

Diebold, F.X. (1986): Modeling the Persistence of Conditional Variances: A Comment. *Econometric Reviews* **5**, 51–56.

Ding Z., Granger, W.J. and Engle, R.F. (1993): A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance* **1**, 83–106.

Duan, J.-C. (1997): Augmented GARCH(p,q) Process and its Diffusion Limit. *Journal of Econometrics* **79**, 97–127.

Engle, R.F. (1982): Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **50**, 987–1007.

Engle, R.F. (1995): *ARCH. Selected Readings*. Oxford University Press.

Engle, R.F. (2004): Risk and Volatility: Econometric Models and Financial Practice. *American Economic Review* **94**, 405–420.

Engle, R.F. and Bollerslev, T. (1986): Modeling the Persistence of Conditional Variances. *Econometric Reviews* **5**, 1–50.

Engle, R.F., Lilien, D.M. and Robins, R.P. (1987): Estimating Time-Varying Risk Premia in the Term Structure: The ARCH-M Model. *Econometrica* **55**, 391–407.

Engle, R.F. and Mezrich, J. (1996): GARCH for Groups. *Risk* **9**, no. 8, 36–40.

Engle, R.F. and Ng, V.K. (1993): Measuring and Testing the Impact of News on Volatility. *Journal of Finance* **48**, 1749–1777.

Friedman, M. (1977): Nobel Lecture: Inflation and Unemployment. *Journal of Political Economy* **85**, 451–472.

Geweke, J. (1986): Modeling the Persistence of Conditional Variances: A Comment. *Econometric Reviews* **5**, 57–61.

Giraitis, L., Kokoszka, P. and Leipus, R. (2000): Stationary ARCH Models: Dependence Structure and Central Limit Theorem. *Econometric Theory* **16**, 3–22.

Giraitis, L., Leipus, R. and Surgailis, D. (2006): Recent Advances in ARCH Modelling. *In: Kirman A. and Teyssiere G. (Eds.): Recent Advances in ARCH Modelling, Long Memory in Economics*. Springer, Berlin.

Giraitis, L., Leipus, R. and Surgailis, D. (2008): ARCH(∞) and long memory properties. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 70–84. Springer, New York.

Glosten, R., Jagannathan, R. and Runkle, D. (1993): On the Relation Between Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance* **48**, 1779–1801.

Gonzales-Rivera, G. (1998): Smooth Transition GARCH Models. *Studies in Nonlinear Dynamics and Econometrics* **3**, 161–178.

Gouriéroux, C. (1996): *ARCH Models and Financial Applications*. Springer, Berlin.

Granger, C.W.J. and Ding, Z. (1995): Some Properties of Absolute Returns. An Alternative Measure of Risk. *Annales d'économie et de statistique* **40**, 67–92.

Gray, S.F. (1996): Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process. *Journal of Financial Economics* **42**, 27–62.

Guégan, D. (1994): *Séries chronologiques non linéaires à temps discret*. Economica, Paris.

Hagerud, G. (1997): A New Non-Linear GARCH Model. *EFI Economic Research Institute, Stockholm*.

Hamilton, J.D. and Susmel, R. (1994): Autoregressive Conditional Heteroskedasticity and Changes in Regime. *Journal of Econometrics* **64**, 307–333.

Hass, M., Mittnik, S. and Paolella, M.S. (2004): A New Approach to Markov-Switching GARCH Models. *Journal of Financial Econometrics* **4**, 493–530.

He, C. (2000): Moments and the Autocorrelation Structure of the Exponential GARCH(p, q) Process. *SSE/EFI Working Paper Series in Economics and Finance, Stockholm School of Economics* **359**.

He, C. and Teräsvirta, T. (1999): Properties of Moments of a Family of GARCH Processes. *Journal of Econometrics* **92**, 173–192.

He, C. and Teräsvirta, T. (1999): Properties of the Autocorrelation Function of Squared Observations for Second Order GARCH Processes under Two Sets of Parameter Constraints. *Journal of Time Series Analysis* **20**, 23–30.

He, C., Teräsvirta, T. and Malmsten, H. (2002): Moment Structure of a Family of First-Order Exponential GARCH Models. *Econometric Theory* **18**, 868–885.

He, C., Silvennoinen, A. and Teräsvirta, T. (2006): Parameterizing Unconditional Skewness in Models for Financial Time Series. *Unpublished paper, Stockholm School of Economics*.

Hentschel, L. (1995): All in the Family. Nesting Symmetric and Asymmetric GARCH Models. *Journal of Financial Economics* **39**, 71–104.

Hong, E.P. (1991): The Autocorrelation Structure for the GARCH-M Process. *Economics Letters* **37**, 129–132.

Kim, S., Shephard, N. and Chib, S. (1998): Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies* **65**, 361–393.

Kim, T.-H. and White, H. (2004): On More Robust Estimation of Skewness and Kurtosis. *Finance Research Letters* **1**, 56–73.

Klaassen, F. (2002): Improving GARCH Volatility Forecasts with Regime-Switching GARCH. *Empirical Economics* **27**, 363–394.

Lamoureux, C.G. and Lastrapes, W.G. (1990): Persistence in Variance, Structural Change and the GARCH Model. *Journal of Business and Economic Statistics* **8**, 225–234.

Lange, T. and Rahbek, A. (2008): An introduction to regime switching time series. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 871–887. Springer, New York.

Lanne, M. and Saikkonen, P. (2005): Nonlinear GARCH Models for Highly Persistent Volatility. *Econometrics Journal* **8**, 251–276.

Lee, J.-H. and Brorsen, B.W. (1997): A Non-Nested Test of GARCH vs. EGARCH Models. *Applied Economics Letters* **4**, 765–768.

Li, C.W. and Li, W.K. (1996): On a Double Threshold Autoregressive Heteroskedasticity Time Series Model. *Journal of Applied Econometrics* **11**, 253–274.

Li, W.K., Ling, S. and McAleer, M. (2002): Recent Theoretical Results for Time Series Models with GARCH Errors. *Journal of Economic Surveys* **16**, 245–269.

Lindner, A.M. (2008): Stationarity, mixing, distributional properties and moments of GARCH(p,q)–processes. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 43–69. Springer, New York.

Linton, O. (2008): Semiparametric and nonparametric ARCH modelling. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handook of Financial Time Series*, 156–167. Springer, New York.

Malmsten, H. (2004): Valuating Exponential GARCH Models. *SSE/EFI Working Paper Series in Economics and Finance, Stockholm School of Economics* **564**.

Malmsten, H. and Teräsvirta, T. (2004): Stylized Facts of Financial Time Series and Three Popular Models of Volatility. *SSE/EFI Working Paper Series in Economics and Finance, Stockholm School of Economics* **563**.

Mandelbrot, B. (1963): The Variation of Certain Speculative Prices. *Journal of Business* **36**, 394–419.

Mandelbrot, T. and Taylor, H. (1967): On the Distribution of Stock Price Differences. *Operations Research* **15**, 1057–1062.

Mikosch, T. and Stărică, C. (2004): Nonstationarities in Financial Time Series, the Long-Range Dependence, and the IGARCH Effects. *Review of Economics and Statistics* **66**, 378–390.

Nelson, D.B. (1991): Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* **59**, 347–370.

Nelson, D.B. and Cao, C.Q. (1992): Inequality Constraints in the Univariate GARCH Model. *Journal of Business and Economic Statistics* **10**, 229–235.

Palm, F.C. (1996): GARCH Models of Volatility. *In: Maddala, G.S. and Rao, C.R. (Eds.): Handbook of Statistics 14: Statistical Methods in Finance* **14**, 209–240. Elsevier, Amsterdam.

Pantula, S.G. (1986): Modeling the Persistence of Conditional Variances: A Comment. *Econometric Reviews* **5**, 71–74.

Rabemananjara, R. and Zakoïan, J.M. (1993): Threshold ARCH Models and Asymmetries in Volatility. *Journal of Applied Econometrics* **8**, 31–49.

Rachev, S. and Mittnik, S. (2000): *Stable Paretian Models in Finance*. Wiley, Chichester.

Rydén, T., Teräsvirta, T. and Åsbrink, S. (1998): Stylized Facts of Daily Return Series and the Hidden Markov Model. *Journal of Applied Econometrics* **13**, 217–244.

Schwert, G.W. (1990): Stock Volatility and the Crash of '87. *Review of Financial Studies* **3**, 77–102.

Sentana, E. (1995): Quadratic ARCH Models. *Review of Economic Studies* **62**, 639–661.

Shephard, N.G. (1996): Statistical Aspects of ARCH and Stochastic Volatility. *In: Cox, D.R., Hinkley, D.V. and Barndorff-Nielsen, O.E. (Eds.): Time Series Models in Econometrics, Finance and Other Fields*, 1–67. Chapman and Hall, London.

Silvennoinen, A. and Teräsvirta, T. (2008): Multivariate GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 201–229. Springer, New York.

Straumann, D. (2004): *Estimation in Conditionally Heteroscedastic Time Series Models*. Springer, New York.

Taylor, S. (1986): *Modelling Financial Time Series*. Wiley, Chichester.

Teräsvirta, T. and Zhao, Z. (2006): Stylized Facts of Return Series, Robust Estimates, and Three Popular Models of Volatility. *Stockholm School of Economics* unpublished paper.

Tong, H. (1990): *Non-Linear Time Series. A Dynamical System Approach*. Oxford University Press.

Zakoïan, J.-M. (1994): Threshold Heteroskedastic Models. *Journal of Economic Dynamics and Control* **18**, 931–955.

Stationarity, Mixing, Distributional Properties and Moments of GARCH(p, q)–Processes

Alexander M. Lindner

Abstract This paper collects some of the well known probabilistic properties of GARCH(p, q) processes. In particular, we address the question of strictly and of weakly stationary solutions. We further investigate moment conditions as well as the strong mixing property of GARCH processes. Some distributional properties such as the tail behaviour and continuity properties of the stationary distribution are also included.

1 Introduction

Since their introduction by Engle (1982), autoregressive conditional heteroskedastic (ARCH) models and their extension by Bollerslev (1986) to generalised ARCH (GARCH) processes, GARCH models have been used widely by practitioners. At a first glance, their structure may seem simple, but their mathematical treatment has turned out to be quite complex. The aim of this article is to collect some probabilistic properties of GARCH processes.

Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a sequence of independent and identically distributed (*i.i.d.*) random variables, and let $p \in \mathbb{N} = \{1, 2, \ldots\}$ and $q \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Further, let $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_{p-1} \geq 0$, $\alpha_p > 0$, $\beta_1, \ldots, \beta_{q-1} \geq 0$ and $\beta_q > 0$ be non-negative parameters. A GARCH(p, q) *process* $(X_t)_{t\in\mathbb{Z}}$ with *volatility process* $(\sigma_t)_{t\in\mathbb{Z}}$ is then a solution to the equations

$$X_t = \sigma_t \varepsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \quad t \in \mathbb{Z}, \tag{2}$$

Alexander M. Lindner
Technische Universität Braunschweig, Institut für Mathematische Stochastik, Pockelsstrasse 14, D-38106 Braunschweig, e-mail: a.lindner@tu-bs.de

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_2,

where the process $(\sigma_t)_{t\in\mathbb{Z}}$ is non-negative. The sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ is referred to as the *driving noise sequence*. GARCH$(p,0)$ processes are called ARCH(p) *processes*. The case of a GARCH$(0,q)$ process is excluded since in that case, the volatility equation (2) decouples from the observed process X_t and the driving noise sequence. Note that in some articles (including the original paper by Bollerslev (1986)) the definition of p and q for GARCH processes is interchanged and the process defined in (1) with volatility given by (2) is referred to as GARCH(q,p) rather than GARCH(p,q).

It is a desirable property that σ_t should depend only on the past innovations $(\varepsilon_{t-h})_{h\in\mathbb{N}}$, i.e. be measurable with respect to the σ-algebra generated by $(\varepsilon_{t-h})_{h\in\mathbb{N}}$. If this condition holds, we shall call the GARCH(p,q) process *causal*. Then X_t is measurable with respect to the σ-algebra $\sigma(\varepsilon_{t-h} : h \in \mathbb{N}_0)$ generated by $(\varepsilon_{t-h})_{h\in\mathbb{N}_0}$. Also, σ_t is independent of $(\varepsilon_{t+h})_{h\in\mathbb{N}_0}$, and X_t is independent of $\sigma(\varepsilon_{t+h} : h \in \mathbb{N})$, for fixed t. Often the requirement of causality is added to the definition of GARCH processes. However, since we shall be mainly interested in strictly stationary solutions which turn out to be automatically causal for GARCH processes, we have dropped the requirement at this point.

The requirement that all the coefficients $\alpha_1,\dots,\alpha_p$ and $\beta_1,\dots,\beta_q$ are non-negative ensures that σ_t^2 is non-negative, so that σ_t can indeed be defined as the square root of σ_t^2. The parameter constraints can be slightly relaxed to allow for some negative parameters, but such that σ_t^2 will still be non-negative, see Nelson and Cao (1992). In the present paper, we shall however always assume non-negative coefficients.

The paper is organized as follows: in Section 2 we collect the criteria under which strictly stationary and weakly stationary solutions to the GARCH equations exist. The ARCH(∞) representation for GARCH processes is given in Section 3. In Section 4, we focus on conditions ensuring finiteness of moments, and give the autocorrelation function of the squared observations. Section 5 is concerned with the strong mixing property and an application to the limit behaviour of the sample autocorrelation function when sufficiently high moments exist. In Section 6 we shortly mention the tail behaviour of stationary solutions and their continuity properties. GARCH processes indexed by the integers are addressed in Section 7. Finally, some concluding remarks are made in Section 8.

For many of the results presented in this paper, it was tried to give at least a short sketch of the proof, following often the original articles, or the exposition given by Straumann (2005).

2 Stationary Solutions

Recall that a sequence $(Y_t)_{t\in\mathbb{Z}}$ of random vectors in $\mathbb{R}^d$ is called *strictly stationary*, if for every $t_1,\dots,t_k \in \mathbb{Z}$, the distribution of $(Y_{t_1+h},\dots,Y_{t_k+h})$

does not depend on h for $h \in \mathbb{N}_0$. When speaking of a *strictly stationary* GARCH(p,q) *process*, we shall mean that the bivariate process $(X_t, \sigma_t)_{t\in\mathbb{N}_0}$ is strictly stationary.

2.1 Strict stationarity of ARCH(1) and GARCH(1,1)

Now suppose that $(p,q) = (1,1)$ or that $(p,q) = (1,0)$, that $(\varepsilon_t)_{t\in\mathbb{Z}}$ is i.i.d., and that $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ satisfy (1), (2). Hence we have a GARCH$(1,1)$/ARCH(1) process, whose volatility process satisfies

$$\sigma_t^2 = \alpha_0 + \beta_1\sigma_{t-1}^2 + \alpha_1\sigma_{t-1}^2\varepsilon_{t-1}^2 = \alpha_0 + (\beta_1 + \alpha_1\varepsilon_{t-1}^2)\sigma_{t-1}^2, \tag{3}$$

where $\beta_1 := 0$ if $q = 0$. Denoting

$$A_t = \beta_1 + \alpha_1\varepsilon_t^2, \quad B_t = \alpha_0, \quad \text{and} \quad Y_t = \sigma_{t+1}^2, \tag{4}$$

it follows that $(Y_t)_{t\in\mathbb{Z}} = (\sigma_{t+1}^2)_{t\in\mathbb{Z}}$ is the solution of the random recurrence equation $Y_t = A_tY_{t-1} + B_t$, where $(A_t, B_t)_{t\in\mathbb{Z}}$ is i.i.d. As we shall see, every strictly stationary solution $(\sigma_t^2)_{t\in\mathbb{Z}}$ of (3) can be expressed as an appropriate function of the driving noise sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$, so that stationarity of $(\sigma_t^2)_{t\in\mathbb{Z}}$ implies stationarity of $(\sigma_t^2, \varepsilon_t)_{t\in\mathbb{Z}}$ and hence of (X_t, σ_t). Thus, the question of existence of strictly stationary solutions of the GARCH$(1,1)$ process can be reduced to the study of strictly stationary solutions of (3). Since we will need multivariate random reccurence equations for the treatment of higher order GARCH processes, we give their definition already in $\mathbb{R}^d$. So let $d \in \mathbb{N}$, and suppose $(A_t, B_t)_{t\in\mathbb{Z}}$ is an i.i.d. sequence, where A_t is a $(d \times d)$-random matrix and B_t is a d-dimensional random vector. The difference equation

$$Y_t = A_tY_{t-1} + B_t, \quad t \in \mathbb{Z}, \tag{5}$$

is then called a *random recurrence equation (with i.i.d. coefficients)*, where the solution $(Y_t)_{t\in\mathbb{Z}}$ is a sequence of d-dimensional random vectors. Every such solution then satisfies

$$\begin{aligned} Y_t &= A_tY_{t-1} + B_t \\ &= A_tA_{t-1}Y_{t-2} + A_tB_{t-1} + B_t = \cdots \\ &= \left(\prod_{i=0}^{k} A_{t-i}\right) Y_{t-k-1} + \sum_{i=0}^{k} \left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i} \end{aligned} \tag{6}$$

for all $k \in \mathbb{N}_0$, with the usual convention that $\prod_{j=0}^{-1} A_{t-j} = 1$ for the product over an empty index set. Letting $k \to \infty$, it is reasonable to hope that for a stationary solution, $\lim_{k\to\infty} \left(\prod_{i=0}^{k} A_{t-i}\right) Y_{t-k-1} = 0$ a.s. and

that $\sum_{i=0}^{k}\left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}$ converges almost surely as $k \to \infty$. In the GARCH(1, 1) and ARCH(1) case, this is indeed the case: let A_t, B_t and Y_t as in (4). By (6), we have

$$\sigma_{t+1}^2 = Y_t = \left(\prod_{i=0}^{k} A_{t-i}\right)\sigma_{t-k}^2 + \alpha_0 \sum_{i=0}^{k}\prod_{j=0}^{i-1} A_{t-j}.$$

Since this is a sum of non-negative components, it follows that $\sum_{i=0}^{\infty}\prod_{j=0}^{i-1} A_{t-j}$ converges almost surely for each t, and hence that $\prod_{i=0}^{k} A_{t-i}$ converges almost surely to 0 as $k \to \infty$. Hence if $(\sigma_t^2)_{t\in\mathbb{Z}}$ is strictly stationary, then $\left(\prod_{i=0}^{k} A_{t-i}\right)\sigma_{t-k}^2$ converges in distribution and hence in probability to 0 as $k \to \infty$. So in the ARCH(1) and GARCH(1, 1) case, there is at most one strictly stationary solution $(\sigma_t^2)_{t\in\mathbb{Z}} = (Y_{t-1})_{t\in\mathbb{Z}}$, given by

$$Y_t := \sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}, \quad t \in \mathbb{Z}. \tag{7}$$

On the other hand, it is clear that if (7) converges a.s. for some and hence all $t \in \mathbb{Z}$, where $(A_t, B_t)_{t\in\mathbb{Z}}$ are the i.i.d. coefficients of the random recurrence equation (5) in $\mathbb{R}^d$, then Y_t, defined by (7), defines a strictly stationary solution of (5).

We have seen that existence of a strictly stationary GARCH(1, 1)/ARCH(1) process implies almost sure convergence of $\prod_{i=0}^{k} A_{-i}$ to 0 as $k \to \infty$. For the converse, we cite the following result:

Proposition 1 (Goldie and Maller (2000), Theorem 2.1)
Let $d = 1$ and $(A_t, B_t)_{t\in\mathbb{Z}}$ be i.i.d. in $\mathbb{R}\times\mathbb{R}$. Suppose that $P(B_0 = 0) < 1$, $P(A_0 = 0) = 0$, that $\prod_{i=0}^{n} A_{-i}$ converges almost surely to zero as $n \to \infty$, and that

$$\int_{(1,\infty)} \frac{\log q}{T_A(\log q)}\, P_{|B_0|}(dq) < \infty, \tag{8}$$

where $P_{|B_0|}$ denotes the distribution of $|B_0|$ and $T_A(y) := \int_0^y P(|A_0| < e^{-x})\,dx$ for $y \geq 0$. Then $\sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}$ converges almost surely absolutely for every $t \in \mathbb{Z}$.

In the GARCH(1, 1) / ARCH(1) case, we have $B_0 = \alpha_0 > 0$ and (8) clearly holds. Observe that $\sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}$ converges trivially almost surely if $P(A_0 = 0) > 0$, in which case also $\prod_{i=0}^{\infty} A_{t-i} = 0$ a.s. Hence we see that a strictly stationary solution of GARCH(1, 1) / ARCH(1) exists if and only if $\prod_{i=0}^{k} A_{-i}$ converges almost surely to 0 as $k \to \infty$. If $P(A_0 = 0) > 0$ this is clearly the case, so suppose that $\beta_1 > 0$ or that $P(\varepsilon_0^2 > 0) = 1$. Denoting $W_t := \log A_t$, $\prod_{i=0}^{\infty} A_{-i} = 0$ a.s. is then equivalent to the almost

sure divergence to $-\infty$ of the random walk $S_n := \sum_{i=0}^{n} W_{-n}$. If $EW_0^+ < \infty$, then it is well known that $S_n \to -\infty$ if and only if $EW_0^+ < EW_0^- \leq \infty$, i.e. either $EW_0^- = \infty$ or $E|W_0| < \infty$ with $EW_0 < 0$. Furthermore, S_n cannot diverge almost surely to $-\infty$ as $n \to \infty$ if $EW_0^- < EW_0^+ = \infty$. Observe that in the GARCH(1, 1) case we have $\beta_1 > 0$, so that $W_0 \geq \log \beta_1 > -\infty$, hence $EW_0^- < \infty$, and it follows that there exists a strictly stationary solution of the GARCH(1, 1) process if and only if $E \log(\beta_1 + \alpha_1 \varepsilon_0^2) < 0$. In the ARCH(1) case, however, $EW_0^- = \infty$ can happen. If $EW_0^- = \infty$, it is known from Kesten and Maller (1996) and Erickson (1973), that $S_n \to -\infty$ a.s. if and only if

$$\int_{(0,\infty)} \frac{x}{E(W_0^- \wedge x)} \, dP(W_0^+ \leq x) < \infty.$$

With $W_0 = \log \alpha_1 + \log \varepsilon_0^2$, the latter condition can be easily seen to be independent of $\alpha_1 > 0$. Summing up, we have the following characterisation of stationary solutions of the GARCH(1, 1) and ARCH(1) equations. For the GARCH(1, 1) case, and for the ARCH(1) case with $E \log^+(\varepsilon_0^2) < \infty$ this is due to Nelsen (1990). The ARCH(1) case with $E \log^+(\varepsilon_0^2) = \infty$ was added by Klüppelberg et al. (2004). Here, as usual, for a real number x we set $\log^+(x) = \log(\max(1, x))$, so that $\log^+(\varepsilon_0^2) = (\log \varepsilon_0^2)^+$.

Theorem 1 (Nelsen (1990), Theorem 2, Klüppelberg et al. (2004), Theorem 2.1)
(*a*) *The* GARCH(1, 1) *process with* $\alpha_0, \alpha_1, \beta_1 > 0$ *has a strictly stationary solution if and only if*

$$-\infty < E \log(\beta_1 + \alpha_1 \varepsilon_0^2) < 0. \tag{9}$$

This solution is unique, and its squared volatility is given by

$$\sigma_t^2 = \alpha_0 \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} (\beta_1 + \alpha_1 \varepsilon_{t-1-j}^2). \tag{10}$$

(*b*) *The* ARCH(1) *process with* $\beta_1 = 0$ *and* $\alpha_1, \alpha_0 > 0$ *has a strictly stationary solution if and only if one of the following cases occurs:*

(i) $P(\varepsilon_0 = 0) > 0$.
(ii) $E|\log \varepsilon_0^2| < \infty$ *and* $E \log \varepsilon_0^2 < -\log \alpha_1$, *i.e.* (9) *holds.*
(iii) $E(\log \varepsilon_0^2)^+ < \infty$ *and* $E(\log \varepsilon_0^2)^- = \infty$.
(iv) $E(\log \varepsilon_0^2)^+ = E(\log \varepsilon_0^2)^- = \infty$ *and*

$$\int_0^\infty x \left(\int_0^x P(\log \varepsilon_0^2 < -y) \, dy \right)^{-1} dP(\log \varepsilon_0^2 \leq x) < \infty. \tag{11}$$

In each case, the strictly stationary solution is unique, and its squared volatility is given by (10).

Observe that condition (9) depends on ε_0^2, α_1 and β_1, while conditions (i), (iii) and (iv) in the ARCH case depend on ε_0^2 only.

Example 1 (*a*) Suppose that $(\varepsilon_t)_{t\in\mathbb{Z}}$ is i.i.d. with $E\varepsilon_0^2 \in (0,\infty)$, and suppose that either $\beta_1 > 0$ (GARCH(1,1)) or that $E|\log \varepsilon_0^2| < \infty$. Since

$$E\log(\beta_1 + \alpha_1\varepsilon_0^2) \le \log E(\beta_1 + \alpha_1\varepsilon_0^2) = \log(\beta_1 + E(\varepsilon_0^2)\,\alpha_1)$$

by Jensen's inequality, a sufficient condition for a strictly stationary solution to exist is that $E(\varepsilon_0^2)\,\alpha_1 + \beta_1 < 0$. Now suppose that ε_0 is standard normally distributed. If $\beta_1 = 0$, then

$$E\log(\alpha_1\varepsilon_0^2) = \log\alpha_1 + \frac{4}{\sqrt{2\pi}}\int_0^\infty \log(x)e^{-x^2/2}\,dx = \log(\alpha_1) - (C_{EM} + \log(2)),$$

where $C_{EM} := \lim_{N\to\infty}\sum_{n=1}^N \frac{1}{n} - \log(N) \approx 0.57721566$ is the Euler-Mascheroni constant. Hence, the ARCH(1) process with standard normal noise has a strictly stationary solution if and only

$$\alpha_1 < 2\exp(C_{EM}) \approx 3.562.$$

Since $\lim_{\beta_1\downarrow 0} E\log(\beta_1 + \alpha_1\varepsilon_0^2) = E\log(\alpha_1\varepsilon_0^2)$, it follows that for every $\alpha_1 < 2\exp(C_{EM})$ there exists some $\overline{\beta}(\alpha_1) > 0$ such that the GARCH(1,1) process with parameters α_0, α_1 and $\beta_1 \in (0, \overline{\beta}(\alpha_1))$ and standard normal innovations has a strictly stationary solution. In particular, strictly stationary solutions of the GARCH(1,1) process with $\alpha_1 + \beta_1 > 1$ do exist. However, observe that while α_1 may be bigger than 1, $\beta_1 < 1$ is a necessary condition for a strictly stationary solution to exist.
For normal noise, $E(\log(\beta_1 + \alpha_1\varepsilon_0^2))$ can be expressed in terms of confluent and generalised hypergeometric functions, which in turn can be calculated numerically. See Nelsen (1990), Theorem 6, for details.
(*b*) Consider the ARCH(1) process with $\alpha_1 > 0$, and let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be i.i.d. such that the distribution of ε_0 has atoms at $\pm\sqrt{2 - E_2(2)}$ with mass 1/4 each, and an absolutely continuous component with density $f_\varepsilon(x) = (4|x|(\log|x|)^2)^{-1}\mathbf{1}_{(-1/e,1/e)}(x)$. Here, $E_n(x) = \int_1^\infty e^{-xt}/t^n\,dt$ denotes the exponential integral, and it holds $E_2(2) \approx 0.0375$. Since $\int_{-1/e}^{1/e} f_\varepsilon(x)\,dx = \int_{-\infty}^{-1}(2y^2)^{-1}\,dy = 1/2$, f_ε indeed defines a probability distribution. Moreover, since ε_0 is symmetric, we have $E\varepsilon_0 = 0$ and

$$E\varepsilon_0^2 = \frac{1}{2}\int_0^{1/e}\frac{x}{(\log x)^2}\,dx + \frac{1}{2}(2 - E_2(2)) = \frac{1}{2}\int_{-\infty}^{-1}\frac{e^{2y}}{y^2}\,dy + \frac{1}{2}(2 - E_2(2)) = 1.$$

The absolutely continuous component of $\log\varepsilon_0^2$ can be easily seen to have density $x \mapsto (2x^2)^{-1}\mathbf{1}_{(-\infty,-1)}(x)$, so that $E(\log\varepsilon_0^2)^- = \infty$. Since $E(\log\varepsilon_0^2)^+ < \infty$, the ARCH(1) process with $\alpha_1 > 0$ and the given distribution of the $(\varepsilon_t)_{t\in\mathbb{Z}}$ has a unique strictly stationary solution by Case (iii) of the previous

Theorem.
(*c*) Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be i.i.d. with marginal density

$$f_\varepsilon(x) = \begin{cases} (2|x|(\log|x|)^{3/2})^{-1}, & |x| > e, \\ (4|x|(\log|x|)^2)^{-1}, & 0 < |x| < 1/e, \\ 0, & \text{else.} \end{cases}$$

Then the density of $\log\varepsilon_0^2$ is given by

$$f_{\log\varepsilon^2}(x) = \begin{cases} x^{-3/2}, & x > 1, \\ (2x^2)^{-1}, & x < -1, \\ 0, & x \in [-1,1]. \end{cases}$$

We conclude that $E(\log\varepsilon_0^2)^+ = E(\log\varepsilon_0^2)^- = \infty$, and it is easily checked that (11) is satisfied. Hence, a unique strictly stationary solution of the ARCH(1) process with driving noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ exists.

2.2 Strict stationarity of GARCH(p, q)

For higher order GARCH processes, one has to work with multidimensional random recurrence equations. Consider a GARCH(p,q) process $(X_t)_{t\in\mathbb{Z}}$ with volatility $(\sigma_t)_{t\in\mathbb{Z}}$ and driving noise sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$. Let $\widetilde{p} := \max(p,2)$, $\widetilde{q} := \max(q,2)$ and define the random $(\widetilde{p}+\widetilde{q}-1)$–vectors Y_t and B_t by

$$\begin{aligned} Y_t &= (\sigma_{t+1}^2, \ldots, \sigma_{t-\widetilde{p}+2}^2, X_t^2, \ldots, X_{t-\widetilde{q}+2}^2)' \\ \text{and} \quad B_t &= (\alpha_0, 0, \ldots, 0)' \in \mathbb{R}^{\widetilde{p}+\widetilde{q}-1}\ , \end{aligned} \tag{12}$$

respectively. Further, let $\beta_{q+1} = \beta_2 = 0$ if $q \le 1$, and $\alpha_2 = 0$ if $p = 1$, and define the random $(\widetilde{p}+\widetilde{q}-1)\times(\widetilde{p}+\widetilde{q}-1)$-matrix A_t by

$$A_t = \begin{pmatrix} \beta_1+\alpha_1\varepsilon_t^2 & \beta_2 & \cdots & \beta_{\widetilde{q}-1} & \beta_{\widetilde{q}} & \alpha_2 & \cdots & \alpha_{\widetilde{p}-1} & \alpha_{\widetilde{p}} \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ \epsilon_t^2 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{13}$$

These matrices where introduced by Bougerol and Picard (1992a). It is then easy to see that each strictly stationary solution of the GARCH equations

(1), (2) gives rise to a strictly stationary solution of the random recurrence equation (5) with Y_t, B_t and A_t as defined in (12) and (13), and vice versa. Observe that for $p = q = 1$ and for $(p, q) = (1, 0)$, the random recurrence equation with A_t and B_t as in (12) and (13) differs from the one with A_t and B_t as in (4). In fact, the former is a random recurrence equation in $\mathbb{R}^3$, while the latter is one-dimensional.

Strict stationarity of multivariate random recurrence equations is studied in terms of the top Lyapunov exponent. Let $\|\cdot\|$ be any vector norm in $\mathbb{R}^d$. For a matrix $M \in \mathbb{R}^{d\times d}$, the corresponding matrix norm $\|M\|$ is defined by

$$\|M\| := \sup_{x\in\mathbb{R}, x\neq 0} \frac{\|Mx\|}{\|x\|}.$$

Definition 1 Let $(A_n)_{n\in\mathbb{Z}}$ be an i.i.d. sequence of $d \times d$ random matrices, such that $E\log^+\|A_0\| < \infty$. Then the *top Lyapunov exponent* associated with $(A_n)_{n\in\mathbb{Z}}$ is defined by

$$\gamma := \inf_{n\in\mathbb{N}_0} E\left(\frac{1}{n+1}\|A_0A_{-1}\cdots A_{-n}\|\right).$$

Furstenberg and Kesten (1960) showed that

$$\gamma = \lim_{n\to\infty} \frac{1}{n+1}\log\|A_0A_{-1}\cdots A_{-n}\| \tag{14}$$

almost surely, and an inspection of their proof shows that γ is independent of the chosen vector norm (hence matrix norm).

The existence of stationary solutions of random recurrence equations can be described neatly in terms of strict negativity of the associated top Lyapunov exponent. Namely, Bougerol and Picard (1992b) have shown that so called *irreducible* random recurrence equations with i.i.d. coefficients $(A_t, B_t)_{t\in\mathbb{Z}}$, such that $E\log^+\|A_0\| < \infty$ and $E\log^+\|B_0\| < \infty$, admit a nonanticipative strictly stationary solution if and only if the top Lyapunov exponent associated with $(A_t)_{t\in\mathbb{Z}}$ is strictly negative. Here, *nonanticipative* means that Y_t is independent of $(A_{t+h}, B_{t+h})_{h\in\mathbb{N}}$ for each t. For GARCH(p, q) cases, it is easier to exploit the positivity of the coefficients in the matrix A_t rather than to check that the model is irreducible. The result is again due to Bougerol and Picard:

Theorem 2 (Bougerol and Picard (1992a), Theorem 1.3) *Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be an i.i.d. sequence of random variables such that $E(\log\varepsilon_0^2)^+ < \infty$. Let $\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ be* GARCH(p, q) *parameters, and let the $(\widetilde{p}+\widetilde{q}-1)\times(\widetilde{p}+\widetilde{q}-1)$ random matrices A_t as well as the $(\widetilde{p}+\widetilde{q}-1)$–vectors B_t be defined as in* (13) *and* (12)*, respectively. Then the corresponding* GARCH(p, q) *process admits a strictly stationary solution if and only if the top Lyapunov exponent γ associated with the sequence $(A_t)_{t\in\mathbb{Z}}$ is strictly negative. This solution is unique, and the random vector Y_t defined in* (12) *satisfies* (7)*.*

The fact that every strictly stationary solution must be unique and of the form (7) follows with a refined argument similar to the GARCH(1, 1) case, using that every element in the vectors Y_t and in the matrices A_t must be non-negative. In particular this shows that every strictly stationary solution must be causal (the argument here does not require the assumption of finite log-moments). Further, existence of a strictly stationary solution implies $\lim_{k\to\infty} \|A_0 A_{-1} \cdots A_{-k}\| = 0$ a.s. Since $(A_n)_{n\in\mathbb{Z}}$ is i.i.d. and $E \log^+ \|A_0\| < \infty$, this in turn implies strict negativity of the top Lyapunov exponent γ (see Bougerol and Picard (1992b), Lemma 3.4). That $\gamma < 0$ implies convergence of (7) can be seen from the almost sure convergence in (14), which implies

$$\left\| \left(\prod_{j=0}^{k-1} A_{t-j} \right) B_{t-k} \right\| \leq C_t e^{\gamma k/2}$$

for some random variable C_t. Hence, the series (7) converges almost surely and must be strictly stationary. That strict negativity of the top Lyapunov exponent implies convergence of (7) and hence the existence of strictly stationary solutions is true for a much wider class of random recurrence equations, see e.g. Kesten (1973), Vervaat (1979), Brandt (1986) or Bougerol and Picard (1992b).

Due to its importance, we state the observation made after Theorem 2 again explicitly:

Remark 1 A strictly stationary solution to the GARCH equations (1) and (2) is necessarily unique and the corresponding vector Y_t defined in (12) satisfies (7). In particular, every strictly stationary GARCH process is causal.

For matrices, it may be intractable to obtain explicit expressions for the top Lyapunov exponent and hence to check whether it is strictly negative or not. Often, one has to use simulations based on (14) to do that. If the noise sequence has finite variance, however, Bollerslev gave a handy sufficient condition for the GARCH process to have a strictly stationary solution, which is easy to check (part (a) of the following theorem). Bougerol and Picard showed that the boundary values in this condition can still be attained under certain conditions, and they have also given a necessary condition for strictly stationary solutions to exist:

Corollary 1 (Bollerslev (1986), Theorem 1, Bougerol and Picard (1992a), Corollaries 2.2, 2.3)
Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be the driving noise sequence of a GARCH(p, q) *process, and suppose that $0 < E\varepsilon_0^2 < \infty$. Then the following hold:*
(a) If $E(\varepsilon_0^2) \sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1$, then the GARCH$(p, q)$ *process admits a unique strictly stationary solution.*
(b) If $P(\varepsilon_0 = 0) = 0$, ε_0 has unbounded support, $p, q \geq 2$ and $\alpha_1, \ldots, \alpha_p > 0$, $\beta_1, \ldots, \beta_q > 0$, and $E(\varepsilon_0^2) \sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j = 1$, then the GARCH$(p, q)$ *process admits a unique strictly stationary solution.*

(c) *If* $\sum_{j=1}^{q} \beta_j \geq 1$, *then no strictly stationary solution of the* GARCH(p,q) *process exists.*

For the proof of Corollary 1, one may assume that $E\varepsilon_0^2 = 1$. The general result then follows by an easy transformation. If $E\varepsilon_0^2 = 1$, Bougerol and Picard (1992a) prove (b) by showing that the spectral radius $\rho(E(A_0))$ of the matrix $E(A_0)$ is equal to 1. Recall that the *spectral radius* $\rho(C)$ of a square matrix C is defined by

$$\rho(C) = \sup\{|\lambda| : \lambda \text{ eigenvalue of } C\}.$$

Since A_0 is almost surely not bounded, neither has zero columns nor zero rows, and has non-negative entries, it follows from Theorem 2 of Kesten and Spitzer (1984) that $\gamma < \log \rho(E(A_0)) = 0$. The proofs of (a) and (c) are achieved by similar reasoning, using estimates between the top Lyapunov exponent and the spectral radius. In particular, in case (a) one has $\gamma \leq \log \rho(E(A_0)) < 0$.

For real data one often estimates parameters α_i and β_j such that $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j$ is close to one, when assuming noise with variance 1. In analogy to the integrated ARMA (ARIMA) process, Engle and Bollerslev (1986) call GARCH processes for which $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1$ *integrated* GARCH(p,q) processes, or IGARCH(p,q) *processes*, for short. Observe that Corollary 1(b) shows that IGARCH processes may have a strictly stationary solution, unlike ARIMA processes where a unit root problem occurs.

Remark 2 Let ε_0, p, q and $\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_q$ be as in Corollary 1(b). Then there exists $\delta > 0$ such that for all $\widetilde{\alpha}_i \geq 0$, $\widetilde{\beta}_j \geq 0$ with $|\widetilde{\alpha}_i - \alpha_i| < \delta$ $(i = 1, \ldots, p)$ and $|\widetilde{\beta}_j - \beta_j| < \delta$ $(j = 1, \ldots, q)$, the GARCH(p,q) process with parameters α_0, $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_p$, $\widetilde{\beta}_1, \ldots, \widetilde{\beta}_q$ and noise sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ admits a unique strictly stationary solution. In particular, there exist strictly stationary GARCH(p,q) processes for which $E(\varepsilon_0^2)\sum_{i=1}^{p} \widetilde{\alpha}_i + \sum_{j=1}^{q} \widetilde{\beta}_j > 1$. This follows immediately from Definition 1 and Theorem 2, since for the parameters of Corollary 1(b), the top Lyapunov exponent γ is strictly negative.

2.3 Ergodicity

Let $Y = (Y_t)_{t\in\mathbb{Z}}$ be a strictly stationary time series of random vectors in $\mathbb{R}^k$. Then Y can be seen as a random element in $(\mathbb{R}^k)^{\mathbb{Z}}$, equipped with its Borel-σ-algebra $\mathcal{B}((\mathbb{R}^k)^{\mathbb{Z}})$. Let the backshift operator $\Phi_{BS} : (\mathbb{R}^k)^{\mathbb{Z}} \to (\mathbb{R}^k)^{\mathbb{Z}}$ be given by $\Phi_{BS}((z_i)_{i\in\mathbb{Z}}) = (z_{i-1})_{i\in\mathbb{Z}}$. Then the time series $(Y_t)_{t\in\mathbb{Z}}$ is called *ergodic* if $\Phi_{BS}(\Lambda) = \Lambda$ for $\Lambda \in \mathcal{B}((\mathbb{R}^k)^{\mathbb{Z}})$ implies $P(Y \in \Lambda) \in \{0, 1\}$. See e.g. Ash and Gardner (1975) for this and further properties of ergodic time series. In particular, it is known that if $(g_n)_{n\in\mathbb{Z}}$ is a sequence of measurable functions

$g_n : (\mathbb{R}^k)^{\mathbb{Z}} \to \mathbb{R}^d$ such that $g_{n-1} = g_n \circ \Phi_{BS}$ and $Y = (Y_t)_{t\in\mathbb{Z}}$ is strictly stationary and ergodic with values in $\mathbb{R}^k$, then $(g_n(Y))_{n\in\mathbb{Z}}$ is also strictly stationary and ergodic (see e.g. Brandt et al. (1990), Lemma A 1.2.7). Since the sequence $(A_t, B_t)_{t\in\mathbb{Z}}$ is i.i.d. and hence strictly stationary and ergodic for a GARCH process, it follows that every strictly stationary GARCH process is ergodic, since it can be expressed via (7). This is due to Bougerol and Picard (1992a), Theorem 1.3.

2.4 Weak stationarity

Recall that a time series $(Z_t)_{t\in\mathbb{Z}}$ of random vectors in $\mathbb{R}^d$ is called *weakly stationary* or *wide-sense stationary*, if $E\|Z_t\|^2 < \infty$ for all $t \in \mathbb{Z}$, $E(Z_t) \in \mathbb{R}^d$ is independent of $t \in \mathbb{Z}$, and the covariance matrices satisfy

$$\mathrm{Cov}(Z_{t_1+h}, Z_{t_2+h}) = \mathrm{Cov}(Z_{t_1}, Z_{t_2})$$

for all $t_1, t_2, h \in \mathbb{Z}$. Clearly, every strictly stationary sequence which satisfies $E\|Z_0\|^2 < \infty$ is also weakly stationary. For causal GARCH processes, we shall see that the converse is true also, i.e. that every causal weakly stationary GARCH process is also strictly stationary.

Let (X_t, σ_t) be a GARCH process such that σ_t is independent of ε_t, which is in particular satisfied for causal solutions. Then if $P(\varepsilon_0 = 0) < 1$, it follows from (1) and the independence of σ_t and ε_t that for given $r \in (0, \infty)$, $E|X_t|^r < \infty$ if and only if $E|\varepsilon_t|^r < \infty$ and $E\sigma_t^r < \infty$. Suppose $E\varepsilon_0^2 \in (0, \infty)$, and that (X_t, σ_t) is a GARCH(p, q) process such that $E\sigma_t^2 = E\sigma_{t'}^2 < \infty$ for all $t, t' \in \mathbb{Z}$. Then (2) shows that

$$E(\sigma_0^2) = \alpha_0 + \sum_{i=1}^{p} \alpha_i E(\sigma_0^2)\, E(\varepsilon_0^2) + \sum_{j=1}^{q} \beta_j E(\sigma_0^2).$$

Hence we see that a necessary condition for a causal weakly stationary solution to exist is that $E(\varepsilon_0^2) \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$. Now suppose that $(\sigma_t)_{t\in\mathbb{Z}}$ is a causal weakly stationary solution, and for simplicity assume that $E\varepsilon_0^2 = 1$. With Y_t, B_t and A_t as in (12) and (13), Y_t must satisfy (6). Note that then $\sum_{i=0}^{\infty} \left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}$ converges a.s. to the strictly stationary solution by Corollary 1. By (6), this implies that $\left(\prod_{i=0}^{k} A_{t-i}\right) Y_{t-k-1}$ converges almost surely to some finite random variable as $k \to \infty$. If this limit can be seen to be 0, then it follows that the weakly stationary solution must coincide with the strictly stationary. As remarked after Corollary 1, the spectral radius of $E(A_0)$ is less than 1. Hence there is some $N \in \mathbb{N}$ such that $\|(EA_0)^N\| = \|E(A_0 \cdots A_{-N+1})\| < 1$. By causality and weak stationarity, this implies that $E\left(\left(\prod_{i=0}^{k} A_{t-i}\right) Y_{t-k-1}\right)$ converges to 0 as $k \to \infty$, and since

each of the components of $\left(\prod_{i=0}^{k} A_{t-i}\right) Y_{t-k-1}$ is positive, Fatou's lemma shows that its almost sure limit must be 0, so that every causal weakly stationary solution is also strictly stationary. Conversely, if $(Y_t)_{t\in\mathbb{Z}}$ is a strictly stationary solution and $E(\varepsilon_0^2)\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_q < 1$ with $E\varepsilon_0^2 = 1$ for simplicity, it follows from $\|(EA_0)^N\| < 1$ that $\sum_{i=0}^{\infty} E\left(\left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}\right)$ is finite, and since each of its components is positive, this implies that $E\|Y_t\| < \infty$ for the strictly stationary solution. Summing up, we have the following characterisation of causal weakly stationary solutions, which was derived by Bollerslev (1986).

Theorem 3 (Bollerslev (1986), Theorem 1)
Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be such that $E\varepsilon_0^2 < \infty$. Then the GARCH$(p,q)$ *process $(X_t,\sigma_t)_{t\in\mathbb{Z}}$ admits a causal weakly stationary solution if and only $E(\varepsilon_0^2)\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$. In that case, the causal weakly stationary solution is unique and coincides with the unique strictly stationary solution. It holds*

$$E(\sigma_t^2) = \frac{\alpha_0}{1 - E(\varepsilon_0^2)\sum_{i=1}^{p}\alpha_i - \sum_{j=1}^{q}\beta_j}, \quad E(X_t^2) = E(\sigma_t^2)E(\varepsilon_0^2). \tag{15}$$

3 The ARCH(∞) Representation and the Conditional Variance

Often it can be helpful to view a GARCH(p,q) process as an ARCH process of infinite order. In particular, from the ARCH(∞) representation one can easily read off the conditional variance of X_t given its infinite past $(X_s : s < t)$. Originally, Engle (1982) and Bollerslev (1986) defined ARCH and GARCH processes in terms of the conditional variance. Equation (18) below then shows that this property does hold indeed, so that the definition of GARCH processes given here is consistent with the original one of Engle and Bollerslev.

Theorem 4 (Bollerslev (1986), pp. 309–310)
Let $(X_t,\sigma_t)_{t\in\mathbb{Z}}$ be a strictly stationary GARCH(p,q) *process driven by $(\varepsilon_t)_{t\in\mathbb{Z}}$, such that $E\varepsilon_0^2 < \infty$ and $E(\varepsilon_0^2)\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$. Then there is a sequence $(\psi_j)_{j\in\mathbb{N}_0}$ of real constants such that $\psi_0 > 0$, $\psi_j \geq 0$ for all j, $\sum_{j=0}^{\infty}\psi_j < \infty$, and*

$$\sigma_t^2 = \psi_0 + \sum_{i=1}^{\infty}\psi_i X_{t-i}^2. \tag{16}$$

The constants are determined by

$$\psi_0 = \frac{\alpha_0}{1 - \sum_{j=1}^{q} \beta_j},$$
$$\sum_{j=1}^{\infty} \psi_j z^j = \frac{\sum_{i=1}^{p} \alpha_i z^i}{1 - \sum_{j=1}^{q} \beta_j z^j}, \quad z \in \mathbb{C}, \quad |z| \leq 1. \tag{17}$$

In particular, σ_t^2 is measurable with respect to the infinite past $(X_s : s \leq t-1)$, and the conditional expectation and variance of X_t given $(X_s : s < t)$ are given by

$$E(X_t | X_s : s < t) = E(\varepsilon_0)\,\sigma_t \quad \textit{and} \quad V(X_t | X_s : s < t) = V(\varepsilon_0)\,\sigma_t^2, \tag{18}$$

respectively.

For example, if $(\varepsilon_t)_{t\in\mathbb{Z}}$ is i.i.d. standard normal, then conditionally on $(X_s : s < t)$, X_t is $N(0, \sigma_t^2)$ distributed, since σ_t^2 is a Borel function of $(X_s : s < t)$. ARCH(∞) models were introduced in more generality by Robinson (1991). The explicit expression in (16) can be found in Bollerslev (1986) or Nelson and Cao (1992). It can be derived defining

$$S_t := \sigma_t^2 - E(\sigma_t^2), \quad Z_t := X_t^2 - E(X_t^2), \quad t \in \mathbb{Z}. \tag{19}$$

Then (2) is equivalent to

$$S_t - \sum_{j=1}^{q} \beta_j S_{t-j} = \sum_{i=1}^{p} \alpha_i Z_{t-i}. \tag{20}$$

This is an ARMA equation for $(S_t)_{t\in\mathbb{Z}}$ such that $\sup_{t\in\mathbb{Z}} E|Z_t| < \infty$ and $E(S_t) = E(Z_t) = 0$. Since $\sum_{j=1}^{q} \beta_j < 1$, this ARMA equation is causal, and it follows that $S_t = \sum_{j=1}^{\infty} \psi_j Z_{t-j}$ where $(\psi_j)_{j\in\mathbb{N}}$ are given by (17). An easy calculation prevails that $\psi_j \geq 0$, and resubstituting σ_t^2 and X_t^2 in this ARMA equation shows (16). Hence σ_t is measurable with respect to the σ-algebra generated by $(X_s : s < t)$, while ε_t is independent of this σ-algebra by causality. This then implies (18).

In the literature there exist many other examples of ARCH(∞) models apart from GARCH(p, q). For more information and references regarding ARCH(∞) models, see Giraitis et al. (2006) and (2008).

4 Existence of Moments and the Autocovariance Function of the Squared Process

It is important to know whether the stationary solution has moments of higher order. For example, in Theorem 3, we have seen that the strictly stationary solution has finite second moments if and only if $E(\varepsilon_0^2) \sum_{i=1}^{p} \alpha_i +$

$\sum_{j=1}^{q} \beta_j < 1$, and we have given an explicit expression for $E\sigma_t^2$ and EX_t^2. However, one is also interested in conditions ensuring finiteness of moments of higher order, the most important case being finiteness of $E\sigma_t^4$ and EX_t^4. For the GARCH(1, 1) process with normal innovations, a necessary and sufficient condition for such moments to exist has been given by Bollerslev (1986), and extended by He and Teräsvirta (1999b) to general noise sequences. Ling (1999) and Ling and McAleer (2002) give a necessary and sufficient condition for moments of higher order to exist. For ARCH(p) processes, a necessary and sufficient condition for higher order moments to exist was already obtained earlier by Milhøj (1985).

Observe that if $P(\varepsilon_0 = 0) < 1$, then by independence of X_t and σ_t for strictly stationary and hence causal solutions, the m'th moment of $X_t = \sigma_t\varepsilon_t$ exists if and only $E\sigma_t^m < \infty$ and $E|\varepsilon_t|^m < \infty$. Hence we shall only be concerned with moment conditions for σ_t^2. In most cases, ε_t will be a symmetric distribution, so that the odd moments of ε_t and hence X_t will be zero. The main concern is hence on even moments of GARCH processes.

4.1 Moments of ARCH(1) and GARCH(1, 1)

The following theorem gives a complete characterisation when the (possible fractional) moment of a GARCH(1, 1) or ARCH(1) process exists:

Theorem 5 (Bollerslev (1986), Theorem 2, and He and Teräsvirta (1999b), Theorem 1)
Let (X_t, σ_t) be a strictly stationary GARCH(1, 1) *or* ARCH(1) *process as in* (1), (2). *Let $m > 0$. Then the (fractional) m'th moment $E(\sigma_t^{2m})$ of σ_t^2 exists if and only if*

$$E(\beta_1 + \alpha_1\varepsilon_0^2)^m < 1. \tag{21}$$

If m is a positive integer and this condition is satisfied, and $\mu_j := E(\sigma_t^{2j})$ denotes the j'th moment of σ_t^2, then μ_m can be calculated recursively by

$$\mu_m = (1 - E(\beta_1 + \alpha_1\varepsilon_0^2)^m)^{-1} \sum_{j=0}^{m-1} \binom{m}{j} \alpha_0^{m-j} E(\beta_1 + \alpha_1\varepsilon_0^2)^j \mu_j. \tag{22}$$

The $(2m)$'th moment of X_t is given by

$$E(X_t^{2m}) = \mu_m E(\varepsilon_0^{2m}).$$

That condition (21) is necessary and sufficient for finiteness of $E(\sigma_t^{2m})$ ($m \in (0, \infty)$) can be easily seen from representation (10): for if $E(\beta_1 + \alpha_1\varepsilon_0^2)^m < 1$ and $m \in [1, \infty)$, then Minkowski's inequality shows that

$$(E(\sigma_t^{2m}))^{1/m} \leq \alpha_0 \sum_{i=0}^{\infty} (E(\beta_1 + \alpha_1 \varepsilon_0^2)^m)^{i/m} < \infty,$$

and for $m < 1$ one uses similarly $E(U+V)^m \leq EU^m + EV^m$ for positive random variables U, V. Conversely, if $E(\sigma_t^{2m}) < \infty$, then $E\prod_{j=0}^{i-1}(\beta_1 + \alpha_1 \varepsilon_{t-1-j}^2)^m$ must converge to 0 as $i \to \infty$, which can only happen if (21) holds. Finally, if m is an integer and (21) holds, then (22) follows easily by raising (2) to the m'th power and taking expectations.

Example 2 For an integer m, $E(\sigma_t^{2m})$ is finite if and only if $\sum_{j=0}^{m} \binom{m}{j} \beta_1^{m-j} \alpha_1^j E\varepsilon_t^{2j} < 1$. If ε_t is standard normally distributed, this means that

$$\sum_{j=0}^{\infty} \binom{m}{j} \beta_1^{m-j} \alpha_1^j \prod_{i=1}^{j} (2i-1) < 1.$$

For example, the fourth moment of σ_t exists if and only if $\beta_1^2 + 2\beta_1\alpha_1 + 3\alpha_1^2 < 1$.

As an immediate consequence of Theorem 5, one sees that GARCH processes do not have finite moments of all orders if ε_0 has unbounded support, which is a first indication that GARCH processes will generally have heavy tails:

Corollary 2 *Let $(X_t, \sigma_t : t \in \mathbb{Z})$ be a strictly stationary* GARCH$(1,1)$ *or* ARCH(1) *process and assume that $P(\alpha_1 \varepsilon_0^2 + \beta_1 > 1) > 0$. Then there is $r \geq 1$, such that $E\sigma_0^{2r} = E|X_0|^{2r} = \infty$.*

4.2 Moments of GARCH(p, q)

For GARCH processes of higher order, Ling (1999) and Ling and McAleer (2002) give necessary and sufficient conditions for even moments of σ_t to be finite. In order to state their result, we need the notion of the Kronecker product of two matrices. For an $(m \times n)$-matrix $C = (c_{ij})_{i=1,\ldots,m, j=1,\ldots,n}$ and a $(p \times r)$-matrix D, the *Kronecker product* $C \otimes D$ is the $(mp \times nr)$-matrix given by

$$C \otimes D = \begin{pmatrix} c_{11}D & \cdots & c_{1n}D \\ \vdots & \ddots & \vdots \\ c_{m1}D & \cdots & c_{mn}D \end{pmatrix}.$$

See e.g. Lütkepohl (1996) for elementary properties of the Kronecker product. We then have:

Theorem 6 (Ling and McAleer (2002), Theorem 2.1)
Let $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ be a strictly stationary GARCH(p,q) *process as in* (1), (2), *and assume that $\alpha_1 + \beta_1 > 0$. Let A_t be the $(\widetilde{p}+\widetilde{q}-1)\times(\widetilde{p}+\widetilde{q}-1)$ matrix of* (13). *Let $m \in \mathbb{N}$. Then the m'th moment of σ_t^2 is finite if and only if the spectral radius of the matrix $E(A_t^{\otimes m})$ is strictly less than 1.*

Originally, Ling and McAleer (2002) formulated their result in terms of the spectral radius of a matrix corresponding to another state space representation of GARCH processes than the A_t-matrix defined in (13). The proof, however, is quite similar. We shortly sketch the argument:
Suppose that $\rho(E(A_t^{\otimes m})) = \limsup_{n\to\infty} \|(E(A_t^{\otimes m}))^n\|^{1/n} < 1$. Then there is $\lambda \in (0,1)$ such that $\|(E(A_t^{\otimes m}))^n\| \le \lambda^n$ for large enough n, so that the supremum of all elements of $(E(A_t^{\otimes m}))^n$ decreases exponentially as $n \to \infty$. The same is then true for all elements of $(E(A_t^{\otimes m'}))^n$ for every $m' \in \{1,\dots,n\}$. Now take the m'th Kronecker power of the representation (7) for the vector Y_t defined in (12). For example, for $m = 2$, one has (since $B_t = B_{t-i}$ in (12))

$$\begin{aligned} Y_t^{\otimes 2} &= \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty}\left(\left(\prod_{j_1=0}^{i_1-1} A_{t-j_1}\right)B_t\right)\otimes\left(\left(\prod_{j_2=0}^{i_2-1} A_{t-j_2}\right)B_t\right) \\ &= \sum_{i_1=0}^{\infty}\sum_{i_2=i_1}^{\infty}\left(\prod_{j_1=0}^{i_1-1} A_{t-j_1}^{\otimes 2}\right)\left(\prod_{j_2=i_1}^{i_2-1}(\mathrm{Id}\otimes A_{t-j_2})\right)B_t^{\otimes 2} \\ &\quad + \sum_{i_1=1}^{\infty}\sum_{i_2=0}^{i_1-1}\left(\prod_{j_2=0}^{i_2-1} A_{t-j_2}^{\otimes 2}\right)\left(\prod_{j_1=i_2}^{i_1-1}(A_{t-j_1}\otimes \mathrm{Id})\right)B_t^{\otimes 2}, \end{aligned}$$

where Id denotes the $(\widetilde{p}+\widetilde{q}-1)\times(\widetilde{p}+\widetilde{q}-1)$ identity matrix. Taking expectations and using the exponential decay of the elements, which are all non-negative, this then shows that $E(Y_t^{\otimes m})$ is finite, and hence that $E(\sigma_t^{2m}) < \infty$. The converse is established along similar lines: finiteness of $E(\sigma_t^{2m})$ implies finiteness of $E(Y_t^{\otimes m})$. Using the fact that all appearing matrices and vectors have non-negative entries, this then implies finiteness of $\sum_{i=0}^{\infty}(E(A_t^{\otimes m}))^i B_0^{\otimes m}$ as argued by Ling and McAleer (2002), and making use of the assumption $\alpha_1 + \beta_1 > 0$, this can be shown to imply finiteness of $\sum_{i=0}^{\infty}\|(E(A_t^{\otimes m}))^i\|$, showing that $\rho(E(A_t^{\otimes m})) < 1$.

To check whether the spectral radius of the matrix $E(A_t^{\otimes m})$ is less than 1 or not may be tedious or only numerically achievable. A simple sufficient condition for the existence of moments can however be obtained by developing the ARCH(∞) representation (16) into a *Volterra series expansion*, as described by Giraitis et al. (2006) and (2008). Accordingly, a sufficient condition for the m'th moment of σ_t^2 in an ARCH(∞) process to exist is that $\sum_{j=1}^{\infty}\psi_j(E(|\varepsilon_0|^{2m}))^{1/m} < 1$. This was shown by Giraitis et al. (2000) for $m = 2$ and observed to extend to hold for general $m \ge 1$ by Giraitis et al. (2006). With (17), this gives for the GARCH(p,q) process:

Proposition 2 (Giraitis et al. (2006), Theorem 2.1)
Let $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ be a strictly stationary GARCH(p, q) *process as in* (1), (2), *let $m \in [1, \infty)$, and suppose that $0 < E|\varepsilon_0|^{2m} < \infty$. Then*

$$\left(\frac{\sum_{i=1}^{p} \alpha_i}{1 - \sum_{j=1}^{q} \beta_j}\right)^m E|\varepsilon_0|^{2m} < 1$$

is a sufficient condition for $E(\sigma_0^{2m}) < \infty$.

Observe that $0 < E|\varepsilon_0|^{2m} < \infty$ implies $\sum_{j=1}^{q} \beta_j < 1$ by Corollary 1(c), so that the expressions in the condition above are well-defined.

In econometrics, the kurtosis is often seen as an indicator for tail heaviness. Recall that the *kurtosis* K_R of a random variable R with $ER^4 < \infty$ is defined by $K_R = \frac{ER^4}{(ER^2)^2}$. If (X_t, σ_t) is a stationary GARCH process which admits finite fourth moment, then if follows from Jensen's inequality that

$$EX_t^4 = E(\varepsilon_t^4)E(\sigma_t^4) \geq E(\varepsilon_t^4)(E(\sigma_t^2))^2 = K_{\varepsilon_0}(E(X_t^2))^2,$$

so that $K_{X_0} \geq K_{\varepsilon_0}$. This shows that the kurtosis of the stationary solution is always greater or equal than the kurtosis of the driving noise sequence, giving another indication that GARCH processes lead to comparatively heavy tails.

While Theorem 6 gives a necessary and sufficient condition for even moments to exist, it does not give any information about the form of the moment. The most important higher order moment is the fourth moment of σ_t, and an elegant method to determine $E\sigma_t^4$ was developed by Karanasos (1999). To illustrate it, suppose that $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ is a strictly stationary GARCH(p, q) process as in (1), (2), such that $E(\sigma_t^4) < \infty$, and denote

$$w := E\varepsilon_0^2, \;\; v := E\varepsilon_0^4, \;\; f := E\sigma_0^4 \;\text{ and }\; g := E\sigma_0^2 = \frac{\alpha_0}{1 - w\sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j},$$

where we used (15). Then w, v, and g are known and we want to determine f. For $i \in \mathbb{N}$, denote further

$$\lambda_i := E(\sigma_t^2 X_{t-i}^2) \quad \text{and} \quad c_i := E(\sigma_t^2 \sigma_{t-i}^2).$$

Since $E(X_t^2 | \varepsilon_{t-h} : h \in \mathbb{N}) = w\sigma_t^2$, it further holds for $i \in \mathbb{N}$,

$$w\lambda_i = E(X_t^2 X_{t-i}^2), \;\; wc_i = E(X_t^2 \sigma_{t-i}^2), \;\text{ and }\; f = E(X_t^2 \sigma_t^2)/w = E(X_t^4)/v.$$

Then, taking expectations in each of the equations

$$X_t^2\sigma_t^2 = X_t^2\left(\alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2\right),$$
$$\sigma_t^2\sigma_{t-j}^2 = \sigma_{t-j}^2\left(\alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2\right), \quad j = 1,\dots,q,$$
$$\sigma_t^2 X_{t-j}^2 = X_{t-j}^2\left(\alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2\right), \quad j = 1,\dots,p,$$

one obtains

$$wf = \alpha_0 wg + \sum_{i=1}^{p} w\alpha_i\lambda_i + \sum_{i=1}^{q} w\beta_i c_i, \tag{23}$$
$$c_j = \alpha_0 g + (w\alpha_j + \beta_j)f + \sum_{i=1}^{j-1}(w\alpha_{j-i} + \beta_{j-i})c_i + \sum_{i=1}^{p-j}\alpha_{j+i}\lambda_i + \sum_{i=1}^{q-j}\beta_{j+i}c_i, \quad j = 1,\dots,q, \tag{24}$$
$$\lambda_j = \alpha_0 wg + (v\alpha_j + w\beta_j)f + \sum_{i=1}^{j-1}(w\alpha_{j-i} + \beta_{j-i})\lambda_i + \sum_{i=1}^{p-j} w\alpha_{j+i}\lambda_i + \sum_{i=1}^{q-j} w\beta_{j+i}c_i, \quad j = 1,\dots,p, \tag{25}$$

where $\alpha_i = 0$ for $i > p$ and $\beta_i = 0$ for $i > q$. Substituting c_q from (24) and λ_p from (25) into (23), one obtains a system of $(p + q - 1)$ equations for the unknown variables $(f, c_1, \dots, c_{q-1}, \lambda_1, \dots, \lambda_{p-1})$. See Karanasos (1999), Theorem 3.1, for more information. For another approach to obtain necessary conditions for the fourth moment to exist and to obtain its structure, we refer to He and Teräsvirta (1999a), Theorem 1.

4.3 The autocorrelation function of the squares

If the driving noise process of a strictly and weakly stationary GARCH process has expectation $E\varepsilon_0 = 0$, then $EX_t = E(\varepsilon_0)\,E(\sigma_t) = 0$, and for $h \in \mathbb{N}$ it follows from (18) that

$$E(X_tX_{t-h}) = E\,E(X_tX_{t-h}|X_s : s < t) = E(X_{t-h}E(\varepsilon_0)\sigma_t) = 0,$$

so that $(X_t)_{t\in\mathbb{Z}}$ is (weak) White Noise (provided $E\varepsilon_0^2 \neq 0$), i.e. a weakly stationary sequence whose elements are uncorrelated. This uncorrelatedness is however not preserved in the squares of the GARCH process. Rather do

the squares $(X_t^2)_{t\in\mathbb{Z}}$ satisfy an ARMA equation. This was already observed by Bollerslev (1986), (1988). More precisely, we have:

Theorem 7 (Bollerslev (1986), Section 4)
Let $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ be a strictly stationary GARCH(p, q) *process such that $E\sigma_0^4 < \infty$, $E\varepsilon_0^4 < \infty$ and* $\mathrm{Var}(\varepsilon_0^2) > 0$. *Define*

$$u_t := X_t^2 - (E\varepsilon_t^2)\sigma_t^2 = (\varepsilon_t^2 - E(\varepsilon_t^2))\sigma_t^2, \quad t \in \mathbb{Z}. \tag{26}$$

Then $(u_t)_{t\in\mathbb{Z}}$ is a White Noise sequence with mean zero and variance $E(\sigma_0^4)$ $\mathrm{Var}(\varepsilon_0^2)$, *and*

$$S_t := \sigma_t^2 - \frac{\alpha_0}{1 - (E\varepsilon_0^2)\sum_{i=1}^p \alpha_i - \sum_{j=1}^q \beta_j}, \quad t \in \mathbb{Z},$$

and

$$W_t := X_t^2 - \frac{\alpha_0 E\varepsilon_0^2}{1 - (E\varepsilon_0^2)\sum_{i=1}^p \alpha_i - \sum_{j=1}^q \beta_j}, \quad t \in \mathbb{Z},$$

satisfy the causal ARMA$(\max(p, q), p - 1)$ *and causal* ARMA$(\max(p, q), q)$ *equations*

$$S_t - \sum_{i=1}^{\max(p,q)} ((E\varepsilon_0^2)\alpha_i + \beta_i)S_{t-i} = \sum_{i=1}^{p} \alpha_i u_{t-i}, \quad t \in \mathbb{Z},$$

and

$$W_t - \sum_{i=1}^{\max(p,q)} ((E\varepsilon_0^2)\alpha_i + \beta_i)W_{t-i} = u_t - \sum_{j=1}^{q} \beta_j u_{t-j}, \quad t \in \mathbb{Z},$$

respectively. Here, $\alpha_i = 0$ for $i > p$ and $\beta_j = 0$ for $j > q$. In particular, the autocovariance and autocorrelation functions of $(\sigma_t^2)_{t\in\mathbb{Z}}$ and that of $(X_t^2)_{t\in\mathbb{Z}}$ are those of the corresponding ARMA *processes.*

The fact that $(u_t)_{t\in\mathbb{Z}}$ is White Noise follows in complete analogy to the White Noise property of $(X_t)_{t\in\mathbb{Z}}$ by using (18). The ARMA representations then follow by inserting (26) into (2), and they are causal by Theorem 3. Observe that the ARMA equation for $(S_t)_{t\in\mathbb{Z}}$ is actually an ARMA$(\max(p, q), p' - 1)$-equation driven by $(u_{t-p'})_{t\in\mathbb{Z}}$, where $p' := \min\{j \in \{1, \ldots, p\} : \alpha_j \neq 0\}$. For general expressions for the autocoviarance functions of ARMA processes, see Brockwell and Davis (1991), Section 3.3.

5 Strong Mixing

Mixing conditions describe some type of asymptotic independence, which may be helpful in proving limit theorems, e.g. for the sample autocorrelation function or in extreme value theory. There exist many types of mixing conditions, see e.g. Doukhan (1994) for an extensive treatment. For GARCH processes, under weak assumptions one has a very strong notion of mixing, namely β-mixing, which in particular implies strong mixing: let $Y = (Y_t)_{t\in\mathbb{Z}}$ be a strictly stationary time series in $\mathbb{R}^d$, defined on a probability space $(\Omega, \mathcal{F}, P)$. Denote by $\mathcal{F}_{-\infty}^0$ the σ-algebra generated by $(Y_s : s \le 0)$ and by $\mathcal{F}_t^\infty$ the σ-algebra generated by $(Y_s : s \ge t)$, and for $k \in \mathbb{N}$ let

$$\alpha_k^{(SM)} := \sup_{C\in\mathcal{F}_{-\infty}^0, D\in\mathcal{F}_k^\infty} |P(C\cap D) - P(C)P(D)|,$$

$$\beta_k^{(SM)} := \frac{1}{2}\sup\sum_{i=1}^{I}\sum_{j=1}^{J} |P(C_i\cap D_j) - P(C_i)P(D_j)|,$$

where in the definition of $\beta_k^{(SM)}$ the supremum is taken over all pairs of finite partitions $\{C_1,\dots,C_I\}$ and $\{D_1,\dots,D_J\}$ of Ω such that $C_i \in \mathcal{F}_{-\infty}^0$ for each i and $D_j \in \mathcal{F}_k^\infty$ for each j. The constants $\alpha_k^{(SM)}$ and $\beta_k^{(SM)}$ are the *α-mixing coefficients* and *β-mixing coefficients*, respectively, and $(Y_t)_{t\in\mathbb{Z}}$ is called *strongly mixing* (or *α-mixing*) if $\lim_{k\to\infty}\alpha_k^{(SM)} = 0$, and *$\beta$-mixing* (or *absolutely regular*) if $\lim_{k\to\infty}\beta_k^{(SM)} = 0$. It is *strongly mixing with geometric rate* if there are constants $\lambda \in (0,1)$ and c such that $\alpha_k^{(SM)} \le c\lambda^k$ for every k, i.e. if α_k decays at an exponential rate, and *β-mixing with geometric rate* is defined similarly. Since

$$\alpha_k^{(SM)} \le \frac{1}{2}\beta_k^{(SM)},$$

β-mixing implies strong mixing.

Based on results of Mokkadem (1990), Boussama (1998) showed that GARCH processes are beta mixing with geometric rate under weak assumptions, see also Boussama (2006). The proof hereby relies on mixing criteria for Markov chains as developed by Feigin and Tweedie (1985), see also Meyn and Tweedie (1996). Observe that the sequence $Y = (Y_t)_{t\in\mathbb{N}_0}$ of random vectors defined by (12) defines a discrete time Markov chain with state space $\mathbb{R}_+^{\tilde{p}+\tilde{q}-1}$. Boussama (1998) then shows that under suitable assumptions on the noise sequence this Markov chain is *geometrically ergodic*, i.e. there is a constant $\lambda \in (0,1)$ such that

$$\lim_{n\to\infty}\lambda^{-n}\|p_n(y,\cdot) - \pi(\cdot)\|_{TV} = 0.$$

Here, $p_n(y, E)$ for $y \in \mathbb{R}_+^{\tilde{p}+\tilde{q}-1}$ and $E \in \mathcal{B}(\mathbb{R}_+^{\tilde{p}+\tilde{q}-1})$ denotes the n-step transition probability from y to E, i.e.

$$p_n(y, E) = P(Y_n \in E | Y_0 = y),$$

π denotes the initial distribution of Y_0 which is chosen to be the stationary one, and $\| \cdot \|_{TV}$ denotes the total variation norm of measures. Since geometric ergodicity implies β-mixing of $(Y_t)_{t\in\mathbb{Z}}$ with geometric rate, using the causality it can be shown that this in turn implies β-mixing of $(\sigma_t, \varepsilon_t)_{t\in\mathbb{Z}}$ and hence of $(X_t)_{t\in\mathbb{Z}}$. Originally, the results in Boussama (1998) and (2006) are stated under the additional assumption that the noise sequence has finite second moment, but an inspection of the proof shows that it is sufficient to suppose that $E|\varepsilon_0|^s < \infty$ for some $s > 0$. The next Theorem gives the precise statements. See also Basrak et al. (2002), Corollary 3.5, and Mikosch and Straumann (2006), Theorem 4.5 and Proposition 4.10.

Theorem 8 (Boussama (1998), Théorème 3.4.2)
Let $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ be a strictly stationary GARCH(p, q) *process as in* (1), (2), *and suppose the noise sequence is such that ε_0 is absolutely continuous with Lebesgue density being strictly positive in a neighbourhood of zero, and such that there exists some $s \in (0, \infty)$ such that $E|\varepsilon_0|^s < \infty$. Let Y_t be defined as in* (12). *Then $(Y_t)_{t\in\mathbb{Z}}$ is β-mixing with geometric rate. In particular, $(\sigma_t^2)_{t\in\mathbb{Z}}$, $(X_t^2)_{t\in\mathbb{Z}}$ and $(X_t)_{t\in\mathbb{Z}}$ are β-mixing and hence strongly mixing with geometric rate.*

An important application of strong mixing is the asymptotic normality of the sample autocovariance and autocorrelation function, under suitable moment conditions. Recall that the *sample autocovariance function* of a time series $(Z_t)_{t\in\mathbb{Z}}$ based on observations $Z_1, \ldots, Z_n$ is defined by

$$\gamma_{Z,n}(h) := \frac{1}{n} \sum_{t=1}^{n-h} (Z_t - \overline{Z}_n)(Z_{t+h} - \overline{Z}_n), \quad h \in \mathbb{N}_0,$$

where $\overline{Z}_n := \frac{1}{n} \sum_{t=1}^n Z_t$ denotes the sample mean. Similarly, the *sample autocorrelation function* is given by

$$\rho_{Z,n}(h) := \frac{\gamma_{Z,n}(h)}{\gamma_{Z,n}(0)}, \quad h \in \mathbb{N}_0.$$

If now $(Z_t)_{t\in\mathbb{Z}}$ is a strictly stationary strongly mixing time series with geometric rate such that $E|Z_t|^{4+\delta} < \infty$ for some $\delta > 0$, then for each $h \in \mathbb{N}_0$, also $(Z_t Z_{t+h})_{t\in\mathbb{Z}}$ is strongly mixing with geometric rate and $E|Z_t Z_{t+h}|^{2+\delta/2} < \infty$. Then a central limit theorem applies, showing that $\sqrt{n} \sum_{j=1}^n (Z_t Z_{t+h} - E(Z_t Z_{t+h}))$ converges in distribution to a mean zero normal random variable as $n \to \infty$, see e.g. Ibragimov and Linnik (1971), Theorem 18.5.3. More generally, using the Cramér-Wold device, one can show

that the vector $(\sqrt{n}\sum_{j=1}^{n}(Z_tZ_{t+h} - E(Z_tZ_{t+h})))_{h=0,\dots,m}$ converges for every $m \in \mathbb{N}$ to a multivariate normal distribution. Standard arguments as presented in Brockwell and Davis (1991), Section 7.3, then give multivariate asymptotic normality of the sample autocovariance function and hence of the autocorrelation function via the delta method. Applying these results to the GARCH process, we have:

Corollary 3 *Suppose that $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ is a strictly stationary* GARCH *process whose noise sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ is such that ε_0 is absolutely continuous with Lebesgue density being strictly positive in a neighbourhood of zero.*
(a) If there is $\delta > 0$ such that $E|X_t|^{4+\delta} < \infty$, then the sample autocovariance and sample autocorrelation function of $(X_t)_{t\in\mathbb{Z}}$ are asymptotically normal with rate $n^{1/2}$, i.e. for every $m \in \mathbb{N}$ there exists a multivariate normal random vector $(V_0, \dots, V_m)$ with mean zero such that $(\sqrt{n}(\gamma_{n,X}(h) - \gamma_X(h)))_{h=0,\dots,m}$ converges in distribution to $(V_0, \dots, V_m)$ as $n \to \infty$, and $(\sqrt{n}(\rho_{n,X}(h) - \rho_X(h)))_{h=1,\dots,m}$ converges to $(\gamma_X(0))^{-1}(V_h - \rho_X(h)V_0)_{h=1,\dots,m}$ as $n \to \infty$. Here, γ_X and ρ_X denote the true autocovariance and autocorrelation function of $(X_t)_{t\in\mathbb{Z}}$, respectively.
(b) If there is $\delta > 0$ such that $E|X_t|^{8+\delta} < \infty$, then the sample autocovariance and sample autocorrelation functions of $(X_t^2)_{t\in\mathbb{Z}}$ are asymptotically normal with rate $n^{1/2}$.

The above statement can for example be found in Basrak et al. (2002), Theorems 2.13 and 3.6. In practice one often estimates GARCH processes with parameters which are close to IGARCH. Hence the assumption on finiteness of $E|X_t|^{4+\delta}$ is questionable. Indeed, in cases when $EX_t^4 = \infty$, one often gets convergence of the sample autocovariance and autocovariance functions to stable distributions, and the rate of convergence is different from $\sqrt{n}$. For the ARCH(1) case, this was proved by Davis and Mikosch (1998), extended by Mikosch and Stărică (2000) to the GARCH(1, 1) case, and by Basrak et al. (2002) to general GARCH(p, q). See also Davis and Mikosch (2008).

6 Some Distributional Properties

In this section we shortly comment on two other properties of the strictly stationary solution, namely tail behaviour and continuity properties. We have already seen that the kurtosis of a GARCH process is always greater than or equal to the kurtosis of the driving noise sequence. Furthermore, Corollary 2 shows that under any reasonable assumption, a GARCH(1, 1) process will never have moments of all orders. Much more is true. Based on Kesten's (Kesten (1973)) powerful results on the tail behaviour of random recurrence equations (see also Goldie (1991) for a simpler proof in dimension 1), one can deduce that GARCH processes have Pareto tails under weak assumptions. For the ARCH(1) process this was proved by de Haan et al. (1989), for

GARCH(1, 1) by Mikosch and Stărică (2000), and for general GARCH(p, q) processes by Basrak et al. (2002). For a precise statement of these results, we refer to Corollary 1 in the article of Davis and Mikosch (2008) in this volume. For example, for a GARCH(1, 1) process with standard normal noise, it holds for the stationary solutions $(X_t, \sigma_t)_{t\in\mathbb{Z}}$,

$$\lim_{x\to\infty} x^{2\kappa} P(\sigma_0 > x) = c_\sigma,$$

$$\lim_{x\to\infty} x^{2\kappa} P(|X_0| > x) = c_\sigma E(|\varepsilon_0|^{2\kappa}), \quad \lim_{x\to\infty} x^{2\kappa} P(X_0 > x) = \frac{c_\sigma}{2} E(|\varepsilon_0|^{2\kappa}).$$

Here, κ is the unique solution in $(0, \infty)$ to the equation

$$E(\alpha_1 \varepsilon_0^2 + \beta_1)^\kappa = 1,$$

and c_σ is a strictly positive constant.

Regarding continuity properties of stationary solutions of GARCH(p, q) processes, we shall restrict us to the case of GARCH(1, 1) and ARCH(1). Observe that in that case, the strictly stationary solution satisfies the random recurrence equation

$$\sigma_t^2 = \alpha_0 + (\beta_1 + \alpha_1 \varepsilon_{t-1}^2)\sigma_{t-1}^2.$$

Hence if ε_0 is absolutely continuous, so is $\log(\beta_1 + \alpha_1 \varepsilon_{t-1}^2) + \log \sigma_{t-1}^2$ by independence of ε_{t-1} and σ_{t-1}, and we conclude that σ_t^2 must be absolutely continuous. It follows that absolute continuity of ε_0 leads to absolute continuity of the stationary σ_t and hence of the stationary X_t. Excluding the case when ε_0^2 is constant, i.e. when the distribution of σ_t^2 is a Dirac measure, one might wonder whether the stationary distribution σ_t will always be absolutely continuous, regardless whether ε_0 is absolutely continuous or not. For stationary distributions of the related continuous time GARCH processes (COGARCH) introduced by Klüppelberg et al. (2004), this is indeed the case, see Klüppelberg et al. (2006). For the discrete time GARCH(1, 1) process, the author is however unaware of a solution to this question. At least there is the following positive result which is an easy consequence of Theorem 1 of Grincevicius (1980):

Theorem 9 (Grincevicius (1980), Theorem 1)
Let $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ be a strictly stationary GARCH(1, 1) *or* ARCH(1) *process. Then σ_0 is continuous with respect to Lebesgue measure, i.e. cannot have atoms, unless σ_0 is degenerate to a constant, i.e. unless ε_0^2 is constant. Consequently, X_0 does not have atoms unless ε_0^2 is constant or ε_0 has an atom at zero.*

Actually, Grincevičius' result applies to more general situations, but in the GARCH case says that if $\sigma_0^2 = \alpha_0 \sum_{i=1}^{\infty} \prod_{j=1}^{i-1} (\beta_1 + \alpha_1 \varepsilon_{-j}^2)$ has an atom, then there must exist a sequence $(S_n)_{n\in\mathbb{N}_0}$ such that $\prod_{n=1}^{\infty} P(\alpha_0 + (\beta_1 + \alpha_1 \varepsilon_n^2) S_n =$

$S_{n-1}) > 0$. By the i.i.d. assumption on $(\varepsilon_n)_{n\in\mathbb{Z}}$, this can be seen to happen only if ε_0^2 is constant.

7 Models Defined on the Non-Negative Integers

We defined a GARCH process as a time series indexed by the set $\mathbb{Z}$ of integers. This implies that the process has been started in the infinite past. It may seem more natural to work with models which are indexed by the non-negative integers $\mathbb{N}_0$. Let $(\varepsilon_t)_{t\in\mathbb{N}_0}$ be a sequence of i.i.d. random variables, and $p \in \mathbb{N}$, $q \in \mathbb{N}_0$. Further, let $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_{p-1} \geq 0$, $\alpha_p > 0$, $\beta_1, \ldots, \beta_{q-1} \geq 0$ and $\beta_q > 0$ be non-negative parameters. Then by a GARCH(p, q) *process indexed by* $\mathbb{N}_0$, we shall mean a process $(X_t)_{t\in\mathbb{N}_0}$ with *volatility process* $(\sigma_t)_{t\in\mathbb{N}_0}$ which is a solution to the equations

$$X_t = \sigma_t \varepsilon_t, \quad t \in \mathbb{N}_0, \tag{27}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \quad t \geq \max(p, q). \tag{28}$$

The process is called *causal* if additionally σ_t^2 is independent of $(\varepsilon_{t+h})_{h\in\mathbb{N}_0}$ for $t = 0, \ldots, \max(p, q)$. By (28), the latter independence property then easily extends to hold for all $t \in \mathbb{N}_0$.

Recall that every strictly stationary GARCH(p, q) process indexed by $\mathbb{Z}$ is causal by Remark 1. When restricting such a process to $\mathbb{N}_0$, it is clear that we obtain a causal strictly stationary GARCH process indexed by $\mathbb{N}_0$. Conversely, suppose that $(X_t, \sigma_t)_{t\in\mathbb{N}_0}$ is a strictly stationary GARCH process indexed by $\mathbb{N}_0$. Like any strictly stationary process indexed by $\mathbb{N}_0$, it can be extended to a strictly stationary process $(X_t, \sigma_t)_{t\in\mathbb{Z}}$, see Kallenberg (2002), Lemma 10.2. With $\varepsilon_t = X_t/\sigma_t$ for $t < 0$ (observe that $\sigma_t^2 \geq \alpha_0$), one sees that also $(X_t, \sigma_t, \varepsilon_t)_{t\in\mathbb{Z}}$ is strictly stationary. Hence $(\varepsilon_t)_{t\in\mathbb{Z}}$ must be i.i.d., and (27) and (28) continue to hold for $t \in \mathbb{Z}$. Since $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ is strictly stationary, it is causal, and hence so is $(X_t, \sigma_t)_{t\in\mathbb{N}_0}$.

We have seen that there is an easy correspondence between strictly stationary GARCH processes defined on the integers and strictly stationary GARCH processes defined on $\mathbb{N}_0$. This justifies the restriction to GARCH processes indexed by $\mathbb{Z}$, which are mathematically more tractable. Furthermore, strictly stationary GARCH processes indexed by $\mathbb{N}_0$ are automatically causal.

8 Conclusion

In the present paper we have collected some of the mathematical properties of GARCH(p, q) processes $(X_t, \sigma_t)_{t\in\mathbb{Z}}$. The existence of strictly and weakly stationary solutions was characterised, as well as the existence of moments. The GARCH process shares many of the so called *stylised features* observed in financial time series, like a time varying volatility or uncorrelatedness of the observations, while the squared observations are not uncorrelated. The autocorrelation of the squared sequence was in fact seen to be that of an ARMA process. Stationary solutions of GARCH processes have heavy tails, since they are Pareto under weak assumptions. On the other hand, there are some features which are not met by the standard GARCH(p, q) process, such as the leverage effect, to name just one. In order to include these and similar effects, many different GARCH type models have been introduced, such as the EGARCH model by Nelson (1991), or many other models. We refer to the article by Teräsvirta (2008) for further information regarding various extensions of GARCH processes.

Acknowledgement I would like to thank Richard Davis and Thomas Mikosch for careful reading of the paper and many valuable suggestions.

References

Ash, R.B. and Gardner, M.F. (1975): *Topics in Stochastic Processes Vol. 27 of Probability and Mathematical Statistics.* Academic Press, New York.

Basrak, B., Davis, R.A. and Mikosch, T. (2002): Regular variation of GARCH processes. *Stochastic Processes and Their Applications* **99**, 95–115.

Brandt, A. (1986): The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability* **18**, 211–220.

Brandt, A., Franken, P.and Lisek, B. (1990): *Stationary Stochastic Models.* Wiley, Chichester.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods.* 2nd edition. Springer, Berlin Heidelberg New York.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T. (1988): On the correlation structure for the generalized autoregressive conditional heteroskedastic process. *Journal of Time Series Analysis* **9**, 121–131.

Bougerol, P. and Picard, N. (1992a): Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* **52**, 115–127.

Bougerol, P. and Picard, N. (1992b): Strict stationarity of generalized autoregressive process. *The Annals of Probability* **20**, 1714–1730.

Boussama, F. (1998): *Ergodicité, mélange et estimation dans le modelés GARCH.* Ph.D. Thesis, Université 7 Paris.

Boussama, F. (2006): Ergodicité des chaînes de Markov à valeurs dans une variété algébraique: application aux modéles GARCH multivariés. *Comptes Rendus. Mathématique. Académie des Sciences* **343**, 275–278.

Davis, R.A. and Mikosch, T. (1998): Limit theory for the sample ACF of stationary processes with heavy tails with applications to ARCH. *The Annals of Statistics* **26**, 2049–2080.

Davis, R.A. and Mikosch, T. (2008): Extreme value theory for GARCH processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 186–200. Springer, New York.

Doukhan, P. (1994): *Mixing. Properties and Examples. Lecture Notes in Statistics* **85**, Springer, Berlin Heidelberg New York.

Engle, R.F. (1982): Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1008.

Engle, R.F. and Bollerslev, T. (1986): Modelling the persistence of conditional variances. With comments and a reply by the authors. *Econometric Reviews* **5**, 1–87.

Erickson, K.B. (1973): The strong law of large numbers when the mean is undefined. *Transactions of the American Mathematical Society* **185**, 371–381.

Feigin, P.D. and Tweedie, R.L. (1985): Random coefficient autoregressive processes: a Markov chain analysis of stationarity and finiteness of moments. *Journal of Time Series Analysis* **6**, 1–14.

Furstenberg, H. and Kesten, H. (1960): Products of random matrices. *Annals of Mathematical Statistics* **31**, 457–469.

Giraitis, L., Kokoszka, P. and Leipus, R. (2000): Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory* **16**, 3–22.

Giraitis, L., Leipus, R. and Surgailis, D. (2006): Recent advances in ARCH modelling. *In: Teyssière, G. and Kriman, A. (Eds): Long Memory in Economics*, 3–38. Springer, Berlin Heidelberg New York.

Giraitis, L., Leipus, R. and Surgailis, D (2008): ARCH(∞) models and long memory properties. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 70–84. Springer, New York.

Goldie, C.M. (1991): Implicit renwal theory and tails of solutions of random equations. *The Annals of Applied Probobility* **1**, 126–166.

Goldie, C.M. and Maller, R.A. (2000): Stability of perpetuities. *The Annals of Probability* **28**, 1195–1218.

Grincevičius, A.K. (1980): Products of random affine transformations. *Lithuanian Mathematical Journal* **20**, 279–282.

de Haan, L., Resnick, S.I., Rootzén, H. and de Vries, C.G. (1989): Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. *Stochastic Processes and Their Applications* **32**, 213–224.

He, C. and Teräsvirta, T. (1999a): Fourth moment structure of the GARCH(p,q) process. *Econometric Theory* **15**, 824–846.

He, C. and Teräsvirta, T. (1999b): Properties of moments of a family of GARCH processes. *Journal of Econometrics* **92**, 173–192.

Ibragimov, I.A. and Linnik, Yu.V. (1971): *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

Kallenberg, O. (2002): *Foundations of Modern Probability. 2nd edition.* Springer, Berlin Heidelberg New York.

Karanasos, M. (1999): The second moment and the autovariance function of the squared errors of the GARCH model. *Journal of Econometrics* **90**, 63–67.

Kesten, H. (1973): Random difference equations and renewal theory of products of random matrices. *Acta Mathematica* **131**, 207–248.

Kesten, H. and Maller, R. (1996): Two renewal theorems for general random walk tending to infinity. *Probability Theory and Related Fields* **106**, 1–38.

Kesten, H. and Spitzer, F. (1984): Convergence in distribution for products of random matrices. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **67**, 363–386.

Klüppelberg, C., Lindner, A. and Maller, R. (2004): A continuous-time GARCH process driven by a Lévy process: stationary and second order behaviour. *Journal of Applied Probability* **41**, 601–622.

Klüppelberg, C., Lindner, A. and Maller, R. (2006): Continuous time volatility modelling: GARCH versus Ornstein models. *In: Kabanov, Yu., Liptser, R. and Stoyanov, J. (Eds.): From Stochastic Calculus to Mathematical Finance. The Shiryaev Festschrift*, 393–419. Springer, Berlin New York Heidelberg.

Ling, S. (1999): On the probabilistic properties of a double threshold ARMA conditional heteroskedastic model. *Journal of Applied Probability* **36**, 688–705.

Ling, S. and McAleer, M. (2002): Necessary and sufficient moment conditions for the GARCH(r,s) and asymmetric power GARCH(r,s) models. *Econometric Theory* **18**, 722–729.

Lütkepohl, H. (1996): *Handbook of Matrices*. Wiley, Chichester.

Meyn, S.P. and Tweedie, R.L. (1996): *Markov Chains and Stochastic Stability*. 3rd edition. Springer, Berlin Heidelberg New York.

Milhøj, A. (1985): The moment structure of ARCH processes. *Scandinavian Journal of Statistics* **12**, 281–292.

Mikosch, T. and Stărică, C. (2000): Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *The Annals of Statistics* **28**, 1427–1451.

Mikosch, T. and Straumann, D. (2006): Stable limits of martingale transforms with application to the estimation of GARCH parameters. *The Annals of Statistics* **34**, 493–522.

Mokkadem, A. (1990): Propriétés de mélange des processus autorégressifs polynomiaux. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique* **26**, 219–260.

Nelsen, D.B. (1990): Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory* **6**, 318–334.

Nelson, D.B. (1991): Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**, 347–370.

Nelson, D.B. and Cao, C.Q. (1992): Inequality constraints in the univariate GARCH model. *Journal of Business and Economic Statistics* **10**, 229–235.

Robinson, P.M. (1991): Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* **47**, 67–84.

Straumann, D. (2005): *Estimation in Conditionally Heterscedastic Time Series Models. Lecture Notes in Statistics* **181**, Springer, Berlin.

Teräsvirta, T. (2008): An introduction to univariate GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 17–42. Springer, New York.

Vervaat, W. (1979): On a stochastic difference equation and a representation of nonnegative infinitely divisible random variables. *Advances in Applied Probability* **11**, 750–783.

ARCH(∞) Models and Long Memory Properties

Liudas Giraitis, Remigijus Leipus and Donatas Surgailis

Abstract ARCH(∞)-models are a natural nonparametric generalization of the class of GARCH(p, q) models which exhibit a rich covariance structure (in particular, hyperbolic decay of the autocovariance function is possible). We discuss stationarity, long memory properties and the limit behavior of partial sums of ARCH(∞) processes as well as some of their modifications (linear ARCH and bilinear models).

1 Introduction

A random process (usually interpreted as financial log-return series) $(r_t) = (r_t)_{t\in\mathbb{Z}}$ is said to satisfy the ARCH(∞) equations if there exists a sequence of standard (zero mean and unit variance) iid random variables (ε_t) and a deterministic sequence $b_j \geq 0$, $j = 0, 1, \ldots$, such that

$$r_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j r_{t-j}^2, \qquad t \in \mathbb{Z}. \tag{1}$$

Liudas Giraitis
Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS, U.K., e-mail: L.Giraitis@qmul.ac.uk

Remigijus Leipus
Vilnius University, Naugarduko 24, Vilnius 03225, Lithuania, and Institute of Mathematics and Informatics, Akademijos 4, LT-08663 Vilnius, Lithuania, e-mail: Remigijus.Leipus@maf.vu.lt

Donatas Surgailis
Vilnius University, Naugarduko 24, Vilnius 03225, Lithuania, and Institute of Mathematics and Informatics, Akademijos 4, LT-08663 Vilnius, Lithuania, e-mail: sdonatas@ktl.mii.lt

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_3,

Moreover, we always assume that (r_t) is *causal*, that is, for any t, r_t has a representation as a measurable function of the present and past values $\varepsilon_s, s \le t$. The last property implies that ε_t is independent of $r_s, s < t$ and therefore r_t have zero conditional mean and a (typically random) conditional variance σ_t^2:

$$E[r_t|r_s, s < t] = 0, \quad \mathrm{Var}[r_t|r_s, s < t] = \sigma_t^2.$$

The class of ARCH(∞) processes includes the parametric ARCH and GARCH models of (Engle (1982)) and (Bollerslev (1986)) (see review by (Teräsvirta (2008))). For instance, the GARCH(p, q) process

$$r_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_j r_{t-j}^2, \tag{2}$$

can be written as (1) with

$$\sigma_t^2 = (1 - \beta(1))^{-1}\alpha_0 + (1 - \beta(L))^{-1}\alpha(L)r_t^2, \tag{3}$$

where $\alpha(L) = \alpha_1 L + \cdots + \alpha_q L^q$, $\beta(L) = \beta_1 L + \cdots + \beta_p L^p$ and L stands for the back-shift operator, $L^j X_t = X_{t-j}$. Equation (3) yields the ARCH(∞) representation of the GARCH(p, q) model with positive *exponentially* decaying weights b_j defined by the generating function $\alpha(z)/(1 - \beta(z)) = \sum_{i=1}^{\infty} b_i z^i$; $b_0 = (1 - \beta(1))^{-1}\alpha_0$.

The ARCH(∞) process was introduced by Robinson (1991) and later studied in Kokoszka and Leipus (2000), Giraitis et al. (2000) (see also the review papers Giraitis et al. (2006), Berkes et al. (2004)). In contrast to GARCH(p, q), an ARCH(∞) process can have autocovariances $\mathrm{Cov}(r_k^2, r_0^2)$ decaying to zero at the rate $k^{-\gamma}$ with $\gamma > 1$ arbitrarily close to 1. That is, *the squares r_t^2 of an ARCH(∞) process with finite fourth moment have short memory in the sense of absolutely summable autocovariances*. Numerous empirical studies (Dacorogna et al. (1993), Ding et al. (1993), Baillie et al. (1996), Ding and Granger (1996), Breidt et al. (1998), Andersen et al. (2001)) confirm that the sample autocorrelations of absolute powers of returns series and volatilities are non-negligible for very large lags, which is often referred to as the *long memory phenomenon of asset returns*. The last fact can be alternatively explained by structural changes in GARCH or linear models (Granger and Hyung (2004), Mikosch and Stărică (2000) and (2003), Liu (2000), Leipus et al. (2005)). Several stationary ARCH-type models were proposed to capture the long memory and other empirical "stylized facts" of asset returns. The long memory of the squared process, in the sense of power-law decay of the autocovariance function, was rigorously established for some stochastic volatility models (Harvey (1998), Robinson (2001), Robinson and Zaffaroni (1997) and (1998), Surgailis and Viano (2002)), the Linear ARCH (LARCH) model (Giraitis et al. (2000) and (2004)) and the bilinear model (Giraitis and Surgailis (2002)).

A stationary process (X_t) with finite variance is said to have *covariance long memory* if the series $\sum_{t\in\mathbb{Z}} |\mathrm{Cov}(X_0, X_t)|$ diverges; otherwise (X_t) exhibits *covariance short memory*. We also say that a stationary process (X_t) has *distributional long memory* (respectively, *distributional short memory*) if its normalized partial sum process [1] $\Big(A_n^{-1}\sum_{t=1}^{[n\tau]}(X_t - B_n) \ : \ \tau \in [0,1]\Big)$, converges, in the sense of weak convergence of the finite dimensional distributions, as $n \to \infty$, to a random process $(Z(\tau))_{\tau\in[0,1]}$ with *dependent increments* (respectively, to a random process with *independent increments*). Covariance long memory and distributional long memory are related though generally different notions. These and other definitions of long memory (long-range dependence) can be found in Beran (1994), Cox (1984), Giraitis and Surgailis (2002) and other papers.

2 Stationary ARCH(∞) Process

2.1 Volterra representations

A formal recursion of (1) yields the following Volterra series expansion of r_t^2:

$$r_t^2 \ = \ \varepsilon_t^2\sigma_t^2 = \varepsilon_t^2 b_0\Big(1 + \sum_{k=1}^{\infty}\sum_{j_1,\cdots,j_k=1}^{\infty} b_{j_1}\cdots b_{j_k}\varepsilon_{t-j_1}^2\cdots\varepsilon_{t-j_1-\cdots-j_k}^2\Big). \quad (4)$$

By taking the expectation on both sides and using the independence of the ε_t's, one obtains

$$Er_t^2 \ = \ b_0\Big\{1 + \sum_{k=1}^{\infty}\Big(\sum_{j=1}^{\infty} b_j\Big)^k\Big\} \ = \ \frac{b_0}{1 - \sum_{j=1}^{\infty} b_j}.$$

Let $B = \sum_{j=1}^{\infty} b_j$. The condition

$$B < 1 \quad (5)$$

is necessary and sufficient for the existence of a unique stationary solution of (1) with $Er_t^2 < \infty$, see Kokoszka and Leipus (2000), Giraitis et al. (2000), Zaffaroni (2000). In a similar way,

$$\lambda^{1/2}\ B < 1, \qquad \lambda := E\varepsilon_0^4 \quad (6)$$

is sufficient for the existence of a stationary solution (r_t) with finite fourth moment $Er_t^4 < \infty$. However, condition (6) is not necessary for the existence of such a solution. In the case of GARCH(1,1) (see (12) below) condition (6)

[1] $[s]$ denotes the integer part of s.

translates into $\alpha\lambda^{1/2}+\beta<1$, while a fourth order stationary solution to (12) exists under the weaker condition

$$(\alpha+\beta)^2+\alpha^2(\lambda-1)<1, \tag{7}$$

see Karanasos (1999), He and Teräsvirta (1999) or Davis and Mikosch (2008). A relevant sufficient and necessary condition in the general case of ARCH(∞) can be obtained by centering the innovations in the (nonorthogonal) representation (4), i.e. by replacing the ε_j^2's by $\kappa\zeta_j+1=\varepsilon_j^2$, where the standardized $\zeta_j=(\varepsilon_j^2-E\varepsilon_j^2)/\kappa$, $\kappa^2=\mathrm{Var}(\varepsilon_0^2)$ have zero mean and unit variance. This leads to the following *orthogonal* Volterra representation of (r_t^2) (Giraitis and Surgailis (2002)):

$$r_t^2=\mu+\kappa\mu\sum_{k=1}^{\infty}\sum_{s_k<\cdots<s_2<s_1\le t} g_{t-s_1}h_{s_1-s_2}\cdots h_{s_{k-1}-s_k}\zeta_{s_1}\cdots\zeta_{s_k}, \tag{8}$$

where $\mu=Er_t^2=b_0/(1-B)$, $h_j=\kappa g_j, j\ge 1$, and where $g_j, j\ge 0$, are the coefficients of the generating function

$$\sum_{j=0}^{\infty} g_j z^j \;=\; \left(1-\sum_{i=1}^{\infty} b_i z^i\right)^{-1}. \tag{9}$$

Let $H^2=\sum_{j=1}^{\infty}h_j^2$. The series (8) converges in mean square if and only if

$$B<1, \qquad H<1. \tag{10}$$

Condition (10) is sufficient and necessary for the existence of a stationary ARCH(∞) solution with finite fourth moment (Giraitis and Surgailis (2002)). By orthogonality, it also follows that

$$\begin{aligned}\mathrm{Cov}(r_t^2,r_0^2)&=\kappa^2\mu^2\sum_{k=1}^{\infty}\sum_{s_k<\cdots<s_1\le 0} g_{-s_1}g_{t-s_1}h_{s_1-s_2}^2\cdots h_{s_{k-1}-s_k}^2\\ &=\kappa^2\mu^2\sum_{s\le 0} g_s g_{t-s}\sum_{k=1}^{\infty}H^{2(k-1)}\\ &=\frac{\kappa^2\mu^2}{1-H^2}\sum_{s=0}^{\infty}g_s g_{s+t}.\end{aligned} \tag{11}$$

For the GARCH(1,1) process specified as

$$r_t=\varepsilon_t\sigma_t, \quad \sigma_t^2=\alpha_0+\alpha r_{t-1}^2+\beta\sigma_{t-1}^2 \tag{12}$$

condition (10) is equivalent to (7), $g_j=\alpha(\alpha+\beta)^{j-1}$, and (8) becomes

$$r_t^2 = \mu + \mu\kappa\zeta_t\Big(1 + \sum_{k=1}^{\infty}(\alpha\kappa/\gamma)^k \sum_{s_k<\cdots<s_1<t} \gamma^{t-s_k}\zeta_{s_1}\cdots\zeta_{s_k}\Big)$$
$$+ \mu\sum_{k=1}^{\infty}(\alpha\kappa/\gamma)^k \sum_{s_k<\cdots<s_1<t} \gamma^{t-s_k}\zeta_{s_1}\cdots\zeta_{s_k}, \tag{13}$$

where $\gamma = \alpha + \beta$, $\mu = \alpha_0/(1-\gamma)$. Representation (13) yields the covariance function of the squared GARCH(1,1) process as a function of the parameters $\alpha_0, \alpha, \beta, \lambda$, which was first obtained in Teräsvirta (1996). An alternative approach to the problem of the existence of fourth order stationary solution of ARCH(∞) was discussed in Kazakevičius et al. (2004). Doukhan et al. (2006) discuss vector-valued ARCH(∞) process.

2.2 Dependence structure, association, and central limit theorem

Equation (11) can be applied to study summability and decay properties of the covariance function of (r_t^2) (Giraitis and Surgailis (2002)). Summability of the b_j's implies summability of the g_j's in (9) which in turn implies by (11) that

$$\sum_{k=-\infty}^{\infty} \mathrm{Cov}(r_k^2, r_0^2) < \infty. \tag{14}$$

(Note that $\mathrm{Cov}(r_k^2, r_0^2) \geq 0$ for all k, which follows from (11) and also from the associativity property of (r_t^2), see below.) Therefore, *the squares* (r_t^2) *of a stationary solution of ARCH(∞) with finite fourth moment have covariance short memory.* Giraitis et al. (2000), Giraitis and Surgailis (2002) also prove that hyperbolic decay $b_j \sim Cj^{-\gamma}$ with $\gamma > 1$ implies[2]

$$\mathrm{Cov}(r_k^2, r_0^2) \asymp k^{-\gamma}. \tag{15}$$

Thus, even though condition (10) implies (14), it allows for a hyperbolic rate of decay in (15), with $\gamma > 1$ arbitrarily close to 1. The last property is called *intermediate memory* (Brockwell and Davis (1991), p. 465). A class of ARCH(∞) processes with intermediate memory was discussed in Davidson (2004).

Further insight into the dependence properties of (r_t^2) with finite fourth moment can be obtained from the *moving average representation*

[2] $x_k \sim y_k$ means $x_k/y_k \to 1$ while $x_k \asymp y_k$ means that there are positive constants C_1 and C_2 such that $C_1 y_k < x_k < C_2 y_k$ for all k large enough.

$$r_t^2 = Er_t^2 + \sum_{j=0}^{\infty} g_j \nu_{t-j}, \tag{16}$$

where $(g_j)_{j\geq 0}$ are defined in (9) and $\nu_t = \sigma_t^2(\varepsilon_t^2 - E\varepsilon_t^2) = r_t^2 - \sigma_t^2$ are *martingale differences*, satisfying $E\nu_t^2 < \infty, E[\nu_t | r_s, s < t] = 0$. Representation (16) is a direct consequence of (8), from which ν_t can also be expressed as a Volterra series in the standardized variables $\zeta_s, s < t$. Note that (16) yields the same covariance formula as (11). In the literature, (16) is sometimes obtained without sufficient justification, by treating (ν_t) as "innovations" and formally inverting the equation $r_t^2 - \sum_{j=1}^{\infty} b_j r_{t-j}^2 = b_0 + \nu_t$, see the discussion in Davidson (2004). It is important to realize that the definition of (ν_t) *per se* assumes that (r_t) is a causal solution of (1) and the martingale property of (ν_t) implies $E\sigma_t^2 < \infty$, or $B < 1$, thereby excluding the IGARCH case (see below), for which the sum of the coefficients $B = 1$. On the other hand, even if the fourth moment is finite as in (16), the ν_t's are *not* independent, meaning that "higher order" dependence and distributional properties of (16) may be very different from the usual moving average in iid random variables.

Squared ARCH(∞) processes have the important property of association. A stochastic process (X_t) is said to be *associated* (or *positively correlated*) if

$$\mathrm{Cov}(f(X_{t_1}, \ldots, X_{t_n}), g(X_{t_1}, \ldots, X_{t_n})) \geq 0$$

holds for any coordinate nondecreasing functions $f, g : \mathbf{R}^n \to \mathbf{R}$ and any $t_1, \ldots, t_n, n = 1, 2, \ldots$. In particular, the autocovariance function of an associated process is nonnegative. Association is a very strong property, under which uncorrelatedness implies independence similarly to the Gaussian case. A well-known result due to Newman and Wright (1981) says that if (X_t) is strictly stationary, associated, and $\sum_{t\in\mathbf{Z}} \mathrm{Cov}(X_0, X_t) < \infty$ then its partial sums process converges to a standard Brownian motion in the Skorokhod space $D[0,1]$ with the sup-topology. It is well known that independent random variables are associated, and that this property is preserved by coordinate-nondecreasing (nonlinear) transformations. In particular, the squared ARCH(∞) process of (4) is a coordinate-nondecreasing transformation of the iid sequence (ε_t^2): since all b_j's are nonnegative, r_t^2 can only increase if any of the $\varepsilon_s^2, s \leq t$ on the right-hand side of (4) is replaced by a larger quantity. Therefore the *squared ARCH(∞) process (4) is associated*. The same conclusion holds for any nondecreasing function of (r_t^2), in particular, for fractional powers $(|r_t|^\delta)$ with arbitrary $\delta > 0$.

An immediate consequence from (14), the association property and the above mentioned Newman-Wright theorem is the following functional central limit theorem for the squares of ARCH(∞):

$$\left(n^{-1/2} \sum_{t=1}^{[n\tau]} (r_t^2 - Er_t^2) \right)_{\tau\in[0,1]} \to_{D[0,1]} (\sigma W(\tau))_{\tau\in[0,1]}, \tag{17}$$

where $(W(\tau))$ is a standard Brownian motion and σ^2 equals the sum in (14). According to our terminology, *the squares* (r_t^2) *of a stationary solution of ARCH(∞) with finite fourth moment have distributional short memory.* This result is quite surprising given the rather complicated nonlinear structure of ARCH(∞). Giraitis et al. (2000) obtained a similar result by using finite memory approximations to ARCH(∞).

2.3 Infinite variance and integrated ARCH(∞)

One of the surprising features of ARCH equations is the fact that they may admit a stationary solution which may have arbitrarily heavy power tails, even if the iid "shocks" ε_t are light-tailed, e.g. $N(0,1)$ (see Klivečka and Surgailis (2007)). Bougerol and Picard (1992) obtained sufficient and necessary conditions for the existence of a causal stationary solution of GARCH(p,q), possibly with infinite variance, in terms of the top Lyapunov exponent γ of an associated vector stochastic recurrence equation. The GARCH$(1,1)$ case was first discussed by Nelson (1990). In general, γ is not known explicitly in terms of the coefficients of the GARCH(p,q), the only exception being the GARCH(1,1) case where $\gamma = E\log(\alpha\varepsilon^2 + \beta)$ (Nelson (1990)). The tail behavior of a stationary solution of GARCH(p,q) process was discussed in Basrak et al. (2002), Mikosch and Stărică (2000) (see also Davis and Mikosch (2008) and Lindner (2008) in this volume).

Sufficient conditions for the existence of stationary ARCH(∞) processes without moment conditions were obtained in Kazakevičius and Leipus (2002). They observed that the volatility can be written as

$$\sigma_t^2 = b_0\left(1 + \sum_{n=1}^{\infty}\sigma_{t,n}^2\right),$$

the convergence of the series being equivalent to the existence of a stationary solution $r_t = \varepsilon_t\sigma_t$, where

$$\sigma_{t,n}^2 = \sum_{k=1}^{n}\sum_{\substack{i_1,\cdots,i_k\geq 1\\ i_1+\cdots+i_k=n}} b_{i_1}b_{i_2}\cdots b_{i_k}\varepsilon_{t-i_1}^2\varepsilon_{t-i_1-i_2}^2\cdots\varepsilon_{t-i_1-\cdots-i_k}^2, \quad n\geq 1,$$

$\sigma_{t,n}^2 = 0,\ n\leq 0$, satisfy the recurrence equation

$$\sigma_{t,n}^2 = \varepsilon_{t-n}^2\sum_{i=1}^{n} b_i\sigma_{t,n-i}^2, \quad n\geq 1.$$

The stationarity condition of Kazakevičius and Leipus (2002) (which applies also to the case $E\varepsilon_0^2 = \infty$) essentially reduces to the condition

$$R > 1, \tag{18}$$

where

$$R = \left(\limsup \sqrt[n]{\sigma_{0,n}^2}\right)^{-1}$$

constitutes the (nonrandom) convergence radius of the random power series $\sum_{n=1}^{\infty} \sigma_{0,n}^2 z^n$. In the GARCH$(p,q)$ case, $-\log R = \gamma$ coincides with the top Lyapunov exponent and condition (18) translates to the condition $\gamma < 0$ of Bougerol and Picard (1992). Similarly to GARCH case, the convergence radius R cannot be explicitly determined in terms of the coefficients b_j of ARCH(∞) and therefore the above result does not provide a constructive answer to the question as to the existence of an infinite variance stationary solution to ARCH(∞) except for the case $B = 1$, see below. Kazakevičius and Leipus (2002) also proved the uniqueness of the stationary solution under condition (18) and some additional condition on the coefficients b_j which is satisfied, for example, if b_j ultimately decrease monotonically.

An important class of ARCH(∞) processes with infinite variance are *Integrated ARCH(∞)*, or *IARCH(∞)*, defined as a solution to (1) with

$$B = 1 \tag{19}$$

(recall that $E\varepsilon_t^2 = 1$). For the GARCH(p,q) process, condition (19) becomes the unit root condition $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j = 1$. Thus, the IGARCH$(\infty)$ model is a generalization of the *Integrated GARCH(p,q)* model introduced in Engle and Bollerslev (1986) in order to explain the so-called IGARCH effect of return data when the estimated parameters of the GARCH(p,q) sum up to a value close to 1. Bougerol and Picard (1992) proved that in the IGARCH(p,q) case, the Lyapunov exponent $\gamma < 0$ and therefore equation (1) has a stationary solution.

The above result was extended to the IARCH(∞) equation in Kazakevičius and Leipus (2003), under the additional assumption that the coefficients b_j in (1) decay exponentially. The last condition is crucial in the proof of $R > 1$ in Kazakevičius and Leipus (2003), and is also satisfied in the case of the IGARCH(p,q) discussed in Bougerol and Picard (1992). Kazakevičius and Leipus (2003) also proved that if the exponential decay condition of the b_j's is not satisfied (as in the FIGARCH case), then $R = 1$. However, contrary to the GARCH situation, condition (18) is not necessary for the existence of a stationary solution to ARCH(∞); see Giraitis et al. (2006).

An interesting example of an IARCH process is the *FIGARCH process* defined by

$$r_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = b_0 + (1 - (1-L)^d) r_t^2, \tag{20}$$

where $b_0 > 0$ and $(1-L)^d, 0 < d < 1$, is the fractional differencing operator. This model, introduced by Baillie et al. (1996) in order to capture the long memory effect in volatility, allows for a hyperbolic decay of the coefficients b_j which are positive, summable, and satisfy the unit root condition (19).

However, *the FIGARCH equation has no stationary solution with* $Er_t^2 < \infty$. In general the question of the existence of a stationary solution to (20) with infinite variance remains open. But under additional assumptions on the distribution of ε_t (cf. (20)) it has been shown in Douc et al. (2008) that a non-zero stationary solution exists (related arguments were used also in Robinson and Zaffaroni (2006)).

See Giraitis et al. (2000), Kazakevičius and Leipus (2003), Mikosch and Stărică (2000) and (2003), Davidson (2004) for a discussion of the controversies surrounding the FIGARCH.

3 Linear ARCH and Bilinear Model

The *Linear ARCH (LARCH)* model, introduced by Robinson (1991), is defined by the equations

$$r_t = \sigma_t \varepsilon_t, \quad \sigma_t = \alpha + \sum_{j=1}^{\infty} \beta_j r_{t-j}, \tag{21}$$

where (ε_t) is an iid sequence, with zero mean and unit variance, and the coefficients β_j satisfy $B_2 = \left\{ \sum_{j=1}^{\infty} \beta_j^2 \right\}^{1/2} < \infty$. The particular case $r_t = \sigma_t \varepsilon_t, \quad \sigma_t = (1-L)^{-d} r_t$ corresponds to the LARCH equation with FARIMA$(0, d, 0)$ coefficients.

The main advantage of LARCH is that it allows for modelling of long memory as well as some characteristic asymmetries (the "leverage effect"). Both these properties cannot be modeled by the classical ARCH(∞) with finite fourth moment. Condition $B_2 < \infty$ is weaker than the assumption $B < \infty$ for the ARCH(∞) model (1). Neither α nor the β_j's are assumed positive and, unlike (1), σ_t (not σ_t^2), is a linear combination of the past values of r_t, rather than their squares. Note that $\sigma_t^2 = \mathrm{Var}[r_t | r_s, s < t]$ is the conditional variance of the causal (r_t). Contrary to ARCH(∞), in the LARCH model σ_t may be negative or vanish and so it lacks some of the usual volatility interpretation.

Long memory properties of the LARCH model were studied in Giraitis et al. (2000) and (2004). Similarly as in the ARCH(∞) case, it is easy to show that a second order strictly stationary solution (r_t) to (21) exists if and only if

$$B_2 < 1, \tag{22}$$

in which case it can be represented by the convergent orthogonal Volterra series

$$r_t = \sigma_t \varepsilon_t, \quad \sigma_t = \alpha\Big(1 + \sum_{k=1}^{\infty} \sum_{j_1,\cdots,j_k=1}^{\infty} \beta_{j_1} \cdots \beta_{j_k} \varepsilon_{t-j_1} \cdots \varepsilon_{t-j_1-\cdots-j_k}\Big). \quad (23)$$

Whence, or directly from the LARCH equations (21), relation

$$\mathrm{Cov}(\sigma_t, \sigma_0) = \frac{\alpha^2}{1 - B_2^2} \sum_{j=1}^{\infty} \beta_j \beta_{t+j} \quad (24)$$

immediately follows. Note (24) coincides with the covariance of the linear filter with coefficients $\beta_j(\alpha/(1 - B_2^2))^{1/2}$. In particular, if

$$\beta_j \sim c j^{d-1} \quad (j \to \infty) \quad (25)$$

for some $c > 0, 0 < d < 1/2$, the sequence (σ_t) has covariance long memory similar to FARIMA$(0, d, 0)$ models.

The proof of long memory of (r_t^2) is more involved and uses the combinatorial formalism of *diagrams*, see Giraitis et al. (2000). As is shown in this paper, under some additional boundedness condition on B_2 in (22), *the squared LARCH process of (21), (25) exhibits both covariance and distributional long memory*, since

$$\mathrm{Cov}(r_k^2, r_0^2) \sim C k^{2d-1}, \qquad k \to \infty, \quad (26)$$

and the normalized partial sum process of (r_t^2) tends to a fractional Brownian motion with Hurst parameter $H = d + 1/2 \in (1/2, 1)$. Similar results under increasingly stringent bounds on B_2 in (22) were proved in Giraitis et al. (2000) for arbitrary powers $(r_t^k), k = 2, 3, \ldots$. Berkes and Horvàth (2003) and Beran (2006) obtained the limit distribution of sums of general nonlinear functions of LARCH process.

The above results suggest certain similarities between the long memory properties of LARCH and a linear moving average process with coefficients β_j in (25). In fact, the first term of the expansion (23) (corresponding to $k = 1$) is exactly the linear process $\sum_{j=1}^{\infty} \beta_j \varepsilon_{t-j}$, up to constant α. The nonlinearity of the LARCH appears when analyzing the behavior of higher-order multiple sums in (23). It turns out that every term $\sum_{j_1,\cdots,j_k} \beta_{j_1} \cdots \beta_{j_k} \varepsilon_{t-j_1} \cdots \varepsilon_{t-j_k}$ behaves similarly as the first (linear) term and contributes to the asymptotic constant C in (26), although these "contributions decay geometrically" with k.

A natural generalization of the LARCH model is the heteroscedastic bilinear equation (Giraitis and Surgailis (2002)):

$$X_t = c_0 + \sum_{j=1}^{\infty} c_j X_{t-j} + \sigma_t \zeta_t, \qquad \sigma_t = a_0 + \sum_{j=1}^{\infty} a_j X_{t-j}, \quad (27)$$

where (ζ_t) are standard iid, with zero mean and unit variance, and $a_j, c_j, j \geq 0$, are real (not necessary nonnegative) coefficients. Clearly, $a_j = 0,\ j \geq 1$, in (27) gives a (linear) AR(∞) equation, while $c_j = 0,\ j \geq 0$, results in the LARCH equation (21). Moreover, the ARCH(∞) model in (1) also turns out to be a special case of (27), by putting $X_t = r_t^2,\ \zeta_t = (\varepsilon_t^2 - E\varepsilon_0^2)/\sqrt{\mathrm{Var}(\varepsilon_0^2)}$. Equation (27) appears naturally when studying the class of processes with the property that the conditional mean $\mu_t = E[X_t|X_s, s < t]$ is a linear combination of $X_s, s < t$, and the conditional variance $\sigma_t^2 = \mathrm{Var}[X_t|X_s, s < t]$ is the *square of a linear combination of* $X_s, s < t$, as it is in the case of a causal solution (X_t) of (27):

$$\sigma_t^2 = \left(a_0 + \sum_{j=1}^{\infty} a_j X_{t-j}\right)^2, \qquad \mu_t = c_0 + \sum_{j=1}^{\infty} c_j X_{t-j}.$$

The bilinear model (27) in the cases $c_0 \neq 0$ and $c_0 = 0$ has different properties. The first case (to which ARCH(∞) reduces) does not essentially allow for long memory, see Giraitis and Surgailis (2002). The second case reduces to a linear filter of LARCH. Indeed, let $\nu_t = \sigma_t \zeta_t$, then (X_t) in (27) satisfies the autoregressive equation $X_t = \sum_{j=1}^{\infty} c_j X_{t-j} + \nu_t$. By inverting the last equation, one obtains

$$X_t = \sum_{j=0}^{\infty} g_j \nu_{t-j}, \tag{28}$$

where $G(z) = (1 - C(z))^{-1} = \sum_{j=0}^{\infty} g_j z^j,\ C(z) = \sum_{j=1}^{\infty} c_j z^j$. With (28) in mind, equation (27) can be rewritten as the LARCH equation

$$\nu_t = \sigma_t \zeta_t, \qquad \sigma_t = a_0 + \sum_{j=1}^{\infty} h_j \nu_{t-j} \tag{29}$$

with coefficients h_j given by $H(z) = \sum_{j=1}^{\infty} h_j z^j = A(z)G(z)$. Under condition $H^2 = \sum_{j=1}^{\infty} h_j^2 < 1$ (and some additional assumptions on the coefficients a_j, c_j, see Giraitis and Surgailis (2002)), equation (29) has a unique stationary causal solution (ν_t) given by a convergent Volterra series which can be substituted into (28), yielding a corresponding solution of the bilinear equation (27), viz.

$$X_t = a_0 \sum_{k=1}^{\infty} \sum_{s_k < \cdots < s_1 \leq t} g_{t-s_1} h_{s_1 - s_2} \cdots h_{s_{k-1} - s_k} \zeta_{s_1} \cdots \zeta_{s_k}. \tag{30}$$

The fact that the conditional mean $\mu_t = \sum_{j=1}^{\infty} g_j \nu_{t-j}$ and the conditional variance $\sigma_t = a_0 + \sum_{j=1}^{\infty} h_j \nu_{t-j}$ of the stationary process in (30) admit moving average representations in the martingale difference sequence (ν_t) suggests that (μ_t) and/or (σ_t) may exhibit covariance long memory, provided the filter coefficients decay slowly:

$$g_j \sim C_1 j^{d_1-1}, \qquad h_j \sim C_2 j^{d_2-1}, \tag{31}$$

with some $C_i \neq 0$, $0 < d_i < 1/2$, $i = 1, 2$. The above mentioned paper presents concrete examples of generating functions of the form $C(z) = 1 - P_1(z)(1-z)^{d_1}$, $A(z) = P_2(z)(1-z)^{d_1-d_2}$, where $P_i(z)$, $i = 1, 2$ satisfy some root conditions, for which the corresponding $G(z), H(z)$ satisfy (31). Consequently, the process (X_t) in (27) may exhibit *double long memory* (i.e. long memory both in the conditional mean and in the conditional variance).

Let us finally mention, that heteroscedastic models with non-zero conditional mean (combinations of the type ARMA-ARCH) have been studied in the econometrics literature; see e.g. Beran (2006), Ling and Li (1998), Li et al. (2002), Teyssière (1997). The last paper also discusses econometric aspects of double long memory.

References

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2001): The distribution of realized exchange rate volatility. *J. Amer. Statist. Assoc.* **96**, 45–55.

Baillie, R.T., Bollerslev, T. and Mikkelsen, H.O. (1996): Fractionally integrated generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **74**, 3–30.

Baillie, R.T., Chung, C.-F. and Tieslau, M.A. (1996): Analysing inflation by the fractionally integrated ARFIMA-GARCH model. *J. Appl. Econometrics* **11**, 23–40.

Basrak, B., Davis, R.A. and Mikosch, T. (2002): Regular variation of GARCH processes. *Stoch. Process. Appl.* **99**, 95–116.

Beran, J. (1994): *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

Beran, J. (2006): Location estimation for LARCH processes. *J. Mult. Anal.* **97**, 1766–1782.

Berkes, I., Horvàth, L. and Kokoszka, P.S. (2004): Probabilistic and statistical properties of GARCH processes. *Fields Inst. Commun.* **44**, 409–429.

Berkes, I. and Horvàth, L. (2003): Asymptotic results for long memory LARCH sequences. *Ann. Appl. Probab.* **13**, 641–668.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**, 307–327.

Bougerol, P. and Picard, N. (1992): Stationarity of GARCH processes and of some nonnegative time series. *J. Econometrics* **52**, 115–127.

Breidt, F.J., Crato, N. and de Lima, P. (1998): On the detection and estimation of long memory in stochastic volatility. *J. Econometrics* **83**, 325–348.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods*. Springer, New York.

Cox, D.R (1984): Long-range dependence: a review. *In: David, H.A. and David, H.T. (Eds.): Statistics: An Appraisal. Proc. 50th Anniversary Conference*, 55–74. Iowa State University Press.

Dacorogna, M.M., Müller, U.A., Nagler, R.J., Olsen, R.B. and Pictet, O.V. (1993): A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *J. Intern. Money and Finance* **12**, 413–438.

Davidson, J. (2004): Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *J. Business and Economic Statist.* **22**, 16–29.

Davis, R.A. and Mikosch, T. (2008): Extreme value theory for GARCH processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 186–200. Springer, New York.

Ding, Z., Granger, C.W.J. and Engle, R.F. (1993): A long memory property of stock market returns and a new model. *J. Emp. Finance* **1**, 83–106.

Ding, Z. and Granger, C.W.J. (1996): Modeling volatility persistence of speculative returns: a new approach. *J. Econometrics* **73**, 185–215.

Douc, R., Roueff, F. and Soulier, P. (2008): On the existence of some ARCH(∞)–processes. *Stochastic Processes and Their Applications* **118**, 755–761.

Doukhan, P., Teyssière, G. and Winant, P. (2006): An LARCH(∞) vector valued process. *In: Bertail, P., Doukhan, P. and Soulier, P. (Eds.): Dependence in Probability and Statistics. Lecture Notes in Statistics* **187**, 307–320. Springer, New York.

Engle, R.F. (1982): Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1008.

Engle, R.F. and Bollerslev, T. (1986): Modelling the persistence of conditional variances. *Econometric Reviews* **27**, 1–50.

Giraitis, L., Kokoszka, P. and Leipus, R. (2000): Stationary ARCH models: dependence structure and Central Limit Theorem. *Econometric Th.* **16**, 3-22.

Giraitis, L., Robinson, P.M. and Surgailis, D. (2000): A model for long memory conditional heteroskedasticity. *Ann. Appl. Probab.* **10**, 1002–1024.

Giraitis, L., Leipus, R., Robinson, P.M. and Surgailis, D. (2004): LARCH, leverage and long memory. *J. Financial Econometrics* **2**, 177–210.

Giraitis, L. and Surgailis, D. (2002): ARCH-type bilinear models with double long memory. *Stoch. Process. Appl.* **100**, 275–300.

Giraitis, L., Leipus, R. and Surgailis, D. (2006): Recent advances in ARCH modelling. *In: Teyssière, G. and Kirman, A. (Eds.): Long Memory in Economics*, 3–38. Springer, New York.

Granger, C.W.J. and Hyung, N. (2004): Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *J. Empirical Finance* **11**, 399–421.

Harvey, A. (1998): Long memory in stochastic volatility. *In: Knight, J. and Satchell, S. (Eds.): Forecasting Volatility in the Financial Markets* 307–320. Butterworth & Heineman.

He, C. and Teräsvirta, T. (1999): Fourth moment structure of the GARCH(p, q) process. *Econometric Th.* **15**, 824–846.

Karanasos, M. (1999): The second moment and the autocovariance function of the squared errors of the GARCH model. *J. Econometrics* **90**, 63–76.

Kazakevičius, V. and Leipus, R. (2002): On stationarity in the ARCH(∞) model. *Econometric Th.* **18**, 1–16.

Kazakevičius, V. and Leipus, R. (2003): A new theorem on existence of invariant distributions with applications to ARCH processes. *J. Appl. Probab.* **40**, 147–162.

Kazakevičius, V., Leipus, R. and Viano, M.-C. (2004): Stability of random coefficient autoregressive conditionally heteroskedastic models and aggregation schemes. *J. Econometrics* **120**, 139–158.

Klivečka, A. and Surgailis, D. (2007): GARCH(1,1) process can have arbitrarily heavy power tails. *Lithuan. Math. J.* **47**, 196–210.

Kokoszka, P. and Leipus, R. (2000): Change-point estimation in ARCH models. *Bernoulli* **6**, 513–539.

Leipus, R., Paulauskas, V. and Surgailis, D. (2005): Renewal regime switching and stable limit laws. *J. Econometrics* **129**, 299–327.

Li, W.K., Ling, S. and McAleer, M. (2002): Recent theoretical results for time series models with GARCH errors. *J. Economic Surv.* **16**, 245–269.

Lindner, A.M. (2008): Stationarity, mixing, distributional properties and moments of GARCH(p,q)-processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 43–69. Springer, New York.

Ling, S. and Li, W.K. (1998): On fractionally integrated autoregressive moving average time series models with conditional heteroskedasticity. *J. Amer. Statist. Assoc.* **92**, 1184–1193.

Liu, M. (2000): Modeling long memory in stock market volatility. *J. Econometrics* **99**, 139–171.

Mikosch, T. and Stărică, C. (2000): Is it really long memory we see in financial returns? *In: Embrechts, P. (Ed.): Extremes and Integrated Risk Management*, 149–168. Risk Books, London.

Mikosch, T. and Stărică, C. (2003): Long-range dependence effects and ARCH modeling. *In: Doukhan, P., Oppenheim, G. and Taqqu, M. S. (Eds.): Theory and Applications of Long-Range Dependence*, 539–459. Birkhäuser, Boston.

Mikosch, T. and Stărică, C. (2000): Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Ann. Statist.* **28**, 1427–1451.

Nelson, D.B. (1990) Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory* **6**, 318–334.

Newman, C.M. and Wright, A.L. (1981): An invariance principle for certain dependent sequences. *Ann. Probab.* **9**, 671–675.

Robinson, P.M. (1991): Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *J. Econometrics* **47**, 67–84.

Robinson, P.M. (2001): The memory of stochastic volatility models. *J. Econometrics* **101**, 195–218.

Robinson, P.M. and Zaffaroni, P. (1997): Modelling nonlinearity and long memory in time series. *Fields Inst. Commun.* **11**, 161–170.

Robinson, P.M. and Zaffaroni, P. (1998): Nonlinear time series with long memory: a model for stochastic volatility. *J. Statist. Plan. Infer.* **68**, 359–371.

Robinson, P.M. and Zaffaroni, P. (2006): Pseudo-maximum likelihood estimation of ARCH(1) models. *Ann. Statist.* **34**, 1049–1074.

Surgailis, D. and Viano, M.-C. (2002): Long memory properties and covariance structure of the EGARCH model. *ESAIM: Probability and Statistics* **6**, 311–329.

Teräsvirta, T. (2008): An introduction to univariate GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 17–42. Springer, New York.

Teräsvirta, T. (1996): Two stylized facts and the GARCH(1,1) model. *Stockholm School of Economics. SSE/EFI Working Paper Series in Economics and Finance* **96**.

Teyssière, G. (1997): Double long-memory financial time series. *Preprint, http://greqam.univ-mrs.fr/pdf/working_papers/1997/97B01S.pdf*

Zaffaroni, P. (2000): Stationarity and memory of ARCH(∞) models. *Econometric Theory* **20**, 147–160.

A Tour in the Asymptotic Theory of GARCH Estimation

Christian Francq and Jean-Michel Zakoïan

Abstract The main estimation methods of the univariate GARCH models are reviewed. A special attention is given to the asymptotic results and the quasi-maximum likelihood method.

1 Introduction

In comparison with other volatility models (e.g. the standard stochastic volatility model) GARCH models are simple to estimate, which has greatly contributed to their popularity. The volatility being a function of the past observations, the likelihood function has an explicit form which makes it easy to handle. A variety of alternative estimation methods can also be considered.

Least squares and quasi-maximum likelihood estimations in ARCH models were considered in the seminal paper Engle (1982). The asymptotic properties of the quasi-maximum likelihood estimator (QMLE) received broad interest in the last 20 years. Pioneering work established consistency and asymptotic normality under strong assumptions on the parameter space and the true parameter value. The problem of finding weak assumptions for the consistency and asymptotic normality of the QMLE in GARCH models has attracted a lot of attention in the statistics literature. The first papers limited their scope to ARCH (see Weiss (1986)) or GARCH(1,1) models (see Lee and Hansen (1994), Lumsdaine (1996)). See Berkes and Horváth (2003), Berkes and Horváth (2004), Berkes et al. (2003), Francq and Zakoïan (2004), Hall and Yao (2003), for recent references on the QMLE of general GARCH(p, q)

Christian Francq
University Lille III, EQUIPPE-GREMARS, Domaine du Pont de bois, BP 60149, 59653 Villeneuve d'Ascq Cedex, France, e-mail: christian.francq@univ-lille3.fr

Jean-Michel Zakoïan University Lille III, EQUIPPE-GREMARS, and CREST, 15 Bd G. Péri, 92245 Malakoff cedex, France, e-mail: zakoian@ensae.fr

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_4,

models. See Straumann (2005) for a recent comprehensive monograph on the estimation of GARCH models.

Numerous GARCH-type models have been introduced and it is simply not possible to consider the estimation of all of them. In this article we limit ourselves to the *standard* GARCH(p,q) model given by the equations

$$\begin{cases} \epsilon_t = \sqrt{h_t}\eta_t \\ h_t = \omega_0 + \sum_{i=1}^{q} \alpha_{0i}\epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_{0j} h_{t-j}, \quad t \in \mathbb{Z} = \{0, \pm 1, \dots\} \end{cases} \tag{1}$$

where

$$\omega_0 > 0, \quad \alpha_{0i} \geq 0 \ \ (i = 1, \dots, q), \quad \beta_{0j} \geq 0 \ \ (j = 1, \dots, p),$$

$$\{\eta_t, t \in \mathbb{Z}\} \quad \text{are iid random variables such that } E\eta_1^2 = 1.$$

We assume that $\epsilon_1, \dots, \epsilon_n$ are observations from the process $(\epsilon_t, t \in \mathbb{Z})$, assumed to be a *strictly stationary, ergodic and nonanticipative* solution of Model (1). Conditions for stationarity are obtained (see Lindner (2008)) from the vector representation

$$\underline{z}_t = \underline{b}_t + A_{0t-1}\underline{z}_{t-1}, \tag{2}$$

where, for $p \geq 2$ and $q \geq 2$,

$$\begin{aligned} \underline{z}_t &= (h_t, \dots, h_{t-p+1}, \epsilon_{t-1}^2, \dots, \epsilon_{t-q+1}^2)' \in \mathbb{R}^{p+q-1}, \\ \underline{b}_t &= (\omega, 0, \dots, 0)' \in \mathbb{R}^{p+q-1}, \\ A_{0t} &= \begin{pmatrix} \tau_t' & \beta_{0p} & \alpha_{02:q-1}' & \alpha_{0q} \\ I_{p-1} & 0 & 0 & 0 \\ \xi_t' & 0 & 0 & 0 \\ 0 & 0 & I_{q-2} & 0 \end{pmatrix}, \end{aligned}$$

with

$$\begin{aligned} \tau_t &= (\beta_{01} + \alpha_{01}\eta_t^2, \beta_{02}, \dots, \beta_{0p-1})' \in \mathbb{R}^{p-1}, \\ \xi_t &= (\eta_t^2, 0, \dots, 0)' \in \mathbb{R}^{p-1}, \\ \alpha_{02:q-1} &= (\alpha_{02}, \dots, \alpha_{0q-1})' \in \mathbb{R}^{q-2}. \end{aligned}$$

A nonanticipative solution (ϵ_t) of model (1) is such that ϵ_t is a measurable function of the $(\eta_{t-i}, i \geq 0)$. Bougerol and Picard (1992) showed that the model has a (unique) strictly stationary non anticipative solution if and only if

$$\gamma(\mathbf{A_0}) < 0,$$

where $\gamma(\mathbf{A_0})$ is the top Lyapunov exponent of the sequence (A_{0t}), that is

$$\gamma(\mathbf{A_0}) = \lim_{t\to\infty} \frac{1}{t} \log \|A_{0t}A_{0t-1}\dots A_{01}\| \quad a.s.$$

where $\|\cdot\|$ denotes any norm on the space of the $(p+q-1)\times(p+q-1)$ matrices. In addition, the strictly stationary solution is ergodic as a measurable function of the $(\eta_{t-i}, i \geq 0)$. Let us mention two important consequences of $\gamma(\mathbf{A_0}) < 0$: (i) $\sum_{j=1}^{p} \beta_{0j} < 1$, and (ii) for some $s > 0$, $E|\epsilon_1|^{2s} < \infty$ (see Lemma 2.3 in Berkes et al. (2003) for the proof). The latter property is crucial to avoid unnecessary moment conditions in the proof of the asymptotic properties of the QMLE.

Throughout the orders p and q are assumed to be known. The vector of parameters is denoted by

$$\theta = (\theta_1, \dots, \theta_{p+q+1})' = (\omega, \alpha_1, \dots, \alpha_q, \beta_1, \dots, \beta_p)'$$

and it belongs to a parameter space $\Theta \subset]0, +\infty[\times[0, \infty[^{p+q}$. The true parameter value $\theta_0 = (\omega_0, \alpha_{01}, \dots, \alpha_{0q}, \beta_{01}, \dots, \beta_{0p})'$ is unknown.

We review the main estimation methods, with special attention to the quasi-maximum likelihood method. The focus will be on asymptotic results rather than on small-sample and numerical issues. We start, in Section 2, by considering the Least-Squares estimator (LSE) for ARCH(q) models, which is simple to compute but requires high moment assumptions. Then we turn to QMLE in Section 3. Section 4 is devoted to efficiency issues. In Section 5 we consider alternative estimators. Finally we discuss in Section 6 the case where some GARCH coefficients are equal to zero, putting the true parameter value on the boundary of the parameter space.

2 Least–Squares Estimation of ARCH Models

In this section we assume $p = 0$. The LSE is obtained from the AR(q) representation for ϵ_t^2:

$$\epsilon_t^2 = \omega_0 + \sum_{i=1}^{q} \alpha_{0i}\epsilon_{t-i}^2 + u_t, \tag{3}$$

where $u_t = \epsilon_t^2 - h_t = (\eta_t^2 - 1)h_t$. The sequence $(u_t, \mathcal{F}_{t-1})_t$ is thus a martingale difference when $E\epsilon_1^2 = Eh_1 < \infty$ and $\mathcal{F}_{t-1}$ denotes the σ-field generated by $\{\eta_u, u < t\}$. Let $\epsilon_0, \dots, \epsilon_{1-q}$ denote arbitrary initial values. Introducing the vector $Z'_{t-1} = \left(1, \epsilon_{t-1}^2, \dots, \epsilon_{t-q}^2\right)$, we get from (3)

$$Y = X\theta_0 + U$$

where

$$X = \begin{pmatrix} Z'_{n-1} \\ \vdots \\ Z'_0 \end{pmatrix}, \quad Y = \begin{pmatrix} \epsilon_n^2 \\ \vdots \\ \epsilon_1^2 \end{pmatrix}, \quad U = \begin{pmatrix} u_n \\ \vdots \\ u_1 \end{pmatrix}.$$

When $X'X$ is non-singular (which can be shown to hold, a.s. for large enough n, under Assumption **A3** given below) the LSE of θ_0 is thus given by:

$$\hat{\theta}_n^{LS} = (\hat{\omega}, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)' = (X'X)^{-1}X'Y.$$

The LSE of $s_0^2 = \mathrm{Var}(u_1)$ is

$$\hat{s}_n^2 = \frac{1}{n-q-1}\|Y - X\hat{\theta}_n^{LS}\|^2 = \frac{1}{n-q-1}\sum_{t=1}^{n}\left\{\epsilon_t^2 - \hat{\omega} - \sum_{i=1}^{q}\hat{\alpha}_i\epsilon_{t-i}^2\right\}^2.$$

It is worth mentioning that the LSE is asymptotically equivalent to the Yule-Walker estimator of the AR(q) model (3) (see Chapter 8 in Brockwell and Davis (1991)) and to the Whittle estimator studied by Giraitis and Robinson (2001), Mikosch and Straumann (2002), Straumann (2005).

The consistency and asymptotic normality of the LSE require some additional assumptions. For identifiability we assume that the distribution of η_t is centered and nondegenerate, i.e. $P(\eta_1^2 = 1) \neq 1$. If $E(\epsilon_1^4) < +\infty$, the LSE can be shown to be strongly consistent (see Bose and Mukherjee (2003)):

$$\hat{\theta}_n^{LS} \to \theta_0, \qquad \hat{s}_n^2 \to s_0^2, \qquad \text{a.s. as} \quad n \to \infty.$$

If, in addition $E(\epsilon_1^8) < +\infty$ the estimator of θ_0 is asymptotically normal (see also Bose and Mukherjee (2003)); more precisely,

$$\sqrt{n}(\hat{\theta}_n^{LS} - \theta_0) \xrightarrow{d} \mathcal{N}\left\{0, \ \ (E\eta_1^4 - 1)A^{-1}BA^{-1}\right\}, \tag{4}$$

where

$$A = E_{\theta_0}(Z_q Z'_q), \qquad B = E_{\theta_0}(h_{q+1}^2 Z_q Z'_q)$$

are non-singular matrices. Note that the vector Z_q does not depend on the initial values $\epsilon_0, \ldots, \epsilon_{1-q}$. Consistent estimators of the matrices A and B are straightforwardly obtained by replacing the theoretical moments by empirical ones.

In the framework of linear regression models, it is well known that for heteroscedastic observations the ordinary LSE is outperformed by the quasi-generalized least squares estimator (QGLSE); see *e.g.* Hamilton (1994) Chapter 8. In our framework the QGLSE is defined by

$$\hat{\theta}_n^{QGLS} = (X'\hat{\Omega}X)^{-1}X'\hat{\Omega}Y,$$

where $\hat{\Omega}$ is a consistent estimator of $\Omega = \mathrm{Diag}(h_n^{-2}, \ldots, h_1^{-2})$. If $\hat{\theta}_n^{LS}$ is computed in a first step, then $\hat{\Omega}$ can be obtained by replacing h_t by

$\hat{\omega} + \sum_{i=1}^{q} \hat{\alpha}_i \epsilon_{t-i}^2$ in Ω. Then the two-stage least squares estimator $\hat{\theta}_n^{QGLS}$ is consistent and asymptotically normal

$$\sqrt{n}(\hat{\theta}_n^{QGLS} - \theta_0) \xrightarrow{d} \mathcal{N}\left\{0,\ (E\eta_1^4 - 1)J^{-1}\right\}, \quad J = E_{\theta_0}(h_{q+1}^{-2} Z_q Z_q'), \tag{5}$$

under the moment assumption $E\epsilon_1^4 < \infty$ when all the ARCH coefficients are strictly positive, and under a slightly stronger moment assumption in the general case; see Bose and Mukherjee (2003), Gouriéroux (1997).

The moment conditions can be made explicit using the vector representation (2). It is shown in Chen and An (1998) that $E(\epsilon_1^4) < +\infty$ if and only if $\rho\{E(A_{01} \otimes A_{01})\} < 1$ where $\otimes$ denotes the Kronecker product and $\rho(A)$ the spectral radius of a square matrix A. More generally, if $E\eta_1^{2m} < \infty$ for some positive integer m then $E(\epsilon_1^{2m}) < +\infty$ if and only if $\rho\left\{E(A_{01}^{\otimes m})\right\} < 1$, where $A_{01}^{\otimes m}$ stands for the kronecker product of A_{01} by itself m times. As can be seen in Table 1, the moment conditions imply strong reduction of the admissible parameter space. It is the main advantage of the QMLE to avoid such restrictions.

Table 1 Conditions for strict stationarity and for the existence of moments of the ARCH(1) model when η_t follows a $\mathcal{N}(0,1)$ or Student distributions normalized in such a way that $E\eta_1^2 = 1$ (St_ν stands for a normalized Student distribution with ν degrees of freedom)

	Strict stationarity	$E\epsilon_t^2 < \infty$	$E\epsilon_t^4 < \infty$	$E\epsilon_t^8 < \infty$
Normal	$\alpha_{01} < 3.562$	$\alpha_{01} < 1$	$\alpha_{01} < 0.577$	$\alpha_{01} < 0.312$
St_3	$\alpha_{01} < 7.389$	$\alpha_{01} < 1$	No	No
St_5	$\alpha_{01} < 4.797$	$\alpha_{01} < 1$	$\alpha_{01} < 0.333$	No
St_9	$\alpha_{01} < 4.082$	$\alpha_{01} < 1$	$\alpha_{01} < 0.488$	$\alpha_{01} < 0.143$

Note that, as in the case of linear regression models, the QGLSE is at least as efficient as the LSE. Indeed, setting $D = h_{q+1} A^{-1} Z_q - h_{q+1}^{-1} J^{-1} Z_q$, the matrix

$$\begin{aligned} E_{\theta_0} DD' &= A^{-1} E_{\theta_0}(h_{q+1}^2 Z_q Z_q') A^{-1} + J^{-1} E_{\theta_0}(h_{q+1}^{-2} Z_q Z_q') J^{-1} \\ &\quad - A^{-1} E_{\theta_0}(Z_q Z_q') J^{-1} - J^{-1} E_{\theta_0}(Z_q Z_q') A^{-1} \\ &= A^{-1} B A^{-1} - J^{-1} \end{aligned}$$

is semi-positive definite.

3 Quasi–Maximum Likelihood Estimation

Gaussian quasi-maximum likelihood estimation has become a very popular method for GARCH models. The basic idea of this approach is to maximize

the likelihood function written under the assumption that the noise (η_t) is Gaussian. Since ϵ_t is then Gaussian conditionally on the past ϵ's and σ's, the likelihood function factorizes under a very tractable form which is maximized to produce the QMLE. The Gaussianity of the noise is inessential for the asymptotic properties of the QMLE. We start by considering the case of *pure* GARCH models, corresponding to the practical situation where a GARCH is estimated on the log-returns.

3.1 Pure GARCH models

Conditionally on initial values $\epsilon_0, \ldots, \epsilon_{1-q}, \tilde{\sigma}_0^2, \ldots, \tilde{\sigma}_{1-p}^2$, let us define recursively

$$\tilde{\sigma}_t^2 = \tilde{\sigma}_t^2(\theta) = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \tilde{\sigma}_{t-j}^2$$

for $t = 1, \ldots, n$. Due to the initial values, the sequence $(\tilde{\sigma}_t^2)$ is not stationary, but can be viewed (see Francq and Zakoïan (2004)) as an approximation of the strictly stationary, ergodic and nonanticipative solution of

$$\sigma_t^2 = \sigma_t^2(\theta) = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \qquad \forall t,$$

under the assumption $\sum_{j=1}^{p} \beta_j < 1$. Note that $\sigma_t^2(\theta_0) = h_t$. The Gaussian quasi-likelihood of the observations $\epsilon_1, \ldots, \epsilon_n$ is the function

$$\tilde{L}_n(\theta) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\tilde{\sigma}_t^2}} \exp\left(-\frac{\epsilon_t^2}{2\tilde{\sigma}_t^2}\right).$$

A QMLE of θ_0 is defined as any measurable solution $\hat{\theta}_n^{QML}$ of

$$\hat{\theta}_n^{QML} = \arg\max_{\theta \in \Theta} \tilde{L}_n(\theta) = \arg\min_{\theta \in \Theta} \tilde{\mathbf{l}}_n(\theta), \tag{6}$$

where

$$\tilde{\mathbf{l}}_n(\theta) = n^{-1} \sum_{t=1}^{n} \tilde{\ell}_t, \qquad \text{and} \quad \tilde{\ell}_t = \tilde{\ell}_t(\theta) = \frac{\epsilon_t^2}{\tilde{\sigma}_t^2} + \log \tilde{\sigma}_t^2.$$

Let $\mathcal{A}_\theta(z) = \sum_{i=1}^{q} \alpha_i z^i$ and $\mathcal{B}_\theta(z) = 1 - \sum_{j=1}^{p} \beta_j z^j$ with $\mathcal{A}_\theta(z) = 0$ if $q = 0$ and $\mathcal{B}_\theta(z) = 1$ if $p = 0$.

The paper Berkes et al. (2003) was the first one where the GARCH(p, q) QMLE was captured in a mathematically rigorous way under weak conditions. Several technical assumptions made in Berkes et al. (2003) were relaxed

by Francq and Zakoïan (2004) and Straumann (2005). The two latter papers show that, under the following assumptions

A1: $\theta_0 \in \Theta$ and Θ is compact,
A2: $\gamma(\mathbf{A_0}) < 0$ and $\forall\ \theta \in \Theta, \quad \sum_{j=1}^{p} \beta_j < 1$,
A3: η_t^2 has a non-degenerate distribution with $E\eta_1^2 = 1$,
A4: if $p > 0$, $\mathcal{A}_{\theta_0}(z)$ and $\mathcal{B}_{\theta_0}(z)$ have no common root, $\mathcal{A}_{\theta_0}(1) \neq 0$, and $\alpha_{0q} + \beta_{0p} \neq 0$,

the QMLE is strongly consistent,

$$\hat{\theta}_n^{QML} \to \theta_0, \qquad \text{a.s. as} \quad n \to \infty. \tag{7}$$

Note that in **A2** the condition for strict stationarity is imposed on the true value of the parameter only. To show where Assumptions **A1**-**A4** are used we present the scheme of proof of (7), the reader being referred to Francq and Zakoïan (2004) and Straumann (2005) for a detailed proof. From the second part of **A2** and the compactness of Θ, we have $\sup_{\theta\in\Theta} \sum_{i=1}^{p} \beta_j < 1$. This inequality is used to show that almost surely $\tilde{\sigma}_t^2(\theta) - \sigma_t^2(\theta) \to 0$ uniformly in $\theta \in \Theta$ as $t \to \infty$, and to show that the initial values do not matter asymptotically:

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left| \tilde{\mathbf{l}}_n(\theta) - \mathbf{l}_n(\theta) \right| = 0 \quad a.s.$$

where $\mathbf{l}_n(\theta)$ is a stationary ergodic sequence defined by replacing $\tilde{\sigma}_t^2(\theta)$ by $\sigma_t^2(\theta)$ in $\tilde{\mathbf{l}}_n(\theta)$. Then the first condition in **A2** and the ergodic theorem show that $\tilde{\mathbf{l}}_n(\theta)$ converges a.s. to the asymptotic criterion $E_{\theta_0}\mathbf{l}_1(\theta)$. For any random variable X, let $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$. Note that $E_{\theta_0}\mathbf{l}_1^+(\theta)$ can be equal to $+\infty$, but $E_{\theta_0}\mathbf{l}_1^-(\theta)$ is always finite (because $\inf_{\theta\in\Theta} \omega > 0$) and $E_{\theta_0}\mathbf{l}_1^+(\theta_0)$ is also finite (because under **A2** we have $E_{\theta_0} h_t^s < \infty$ for some $s > 0$, see Berkes et al. (2003) and Nelson (1990) for a proof of this result). Under the identifiability assumptions **A3** and **A4**, $E_{\theta_0}\mathbf{l}_1(\theta) \geq E_{\theta_0}\mathbf{l}_1(\theta_0)$ with equality if and only if $\theta = \theta_0$. These are the main arguments to show that (7) holds. The rest of the proof does not require additional assumptions.

Notice that the condition $E\eta_1 = 0$ is not required. The assumption that $E\eta_1^2 = 1$ is made for identifiability reasons and is not restrictive provided $E\eta_1^2 < \infty$; see Berkes and Horváth (2003). The identifiability condition **A4** excludes that *all* coefficients α_{0i} be zero when $p > 0$, as well as the overidentification of *both* orders p and q. However, other situations where some coefficients α_{0i} or β_{0j} vanish are allowed. This is worth-noting since it is no longer the case for the asymptotic normality (AN).

Indeed, the main additional assumption required for the AN is that θ_0 belongs to the interior $\overset{\circ}{\Theta}$ of Θ. The case where θ_0 belongs to the boundary of Θ will be considered below. Following Francq and Zakoïan (2004), under Assumptions **A1**-**A4** and

A5: $\theta_0 \in \overset{\circ}{\Theta}$,

A6: $E\eta_1^4 < \infty$,

the QMLE is asymptotically normal; more precisely

$$\sqrt{n}(\hat{\theta}_n^{QML} - \theta_0) \xrightarrow{d} \mathcal{N}\left\{0,\ (E\eta_1^4 - 1)J^{-1}\right\}, \quad J = E_{\theta_0}\frac{1}{\sigma_1^4}\frac{\partial\sigma_1^2}{\partial\theta}\frac{\partial\sigma_1^2}{\partial\theta'}(\theta_0). \quad (8)$$

It is shown in Francq and Zakoïan (2004) that $\sigma_1^{-2}(\partial\sigma_1^2/\partial\theta)$ admits moments of any order. For simplicity we give the arguments in the GARCH(1,1) case. We have $\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 = \omega(1-\beta)^{-1} + \alpha\sum_{i=0}^{\infty}\beta^i\epsilon_{t-i-1}^2$. Thus $\sigma_1^{-2}(\partial\sigma_1^2/\partial\omega)$ and $\sigma_1^{-2}(\partial\sigma_1^2/\partial\alpha)$ are bounded, and therefore admit moments of any order. We have already seen that the strict stationarity condition **A2** implies the existence of some $s \in (0,1)$ such that $E_{\theta_0}|\epsilon_1|^{2s} < \infty$. Using $\partial\sigma_t^2/\partial\beta_j = \sum_{k=1}^{\infty} k\beta^{k-1}(\omega + \alpha\epsilon_{t-1-k}^2)$, $\sigma_t^2 \geq \omega + \beta^k(\omega + \alpha\epsilon_{t-1-k}^2)$, and the elementary inequality $x/(1+x) \leq x^s$ for all $x \geq 0$, we obtain for any $d > 0$

$$\begin{aligned}
\left\|\frac{1}{\sigma_t^2}\frac{\partial\sigma_t^2}{\partial\beta}\right\|_d &\leq \left\|\frac{1}{\beta}\sum_{k=1}^{\infty}\frac{k\beta^k\left(\omega + \alpha\epsilon_{t-k-1}^2\right)}{\omega + \beta^k\left(\omega + \alpha\epsilon_{t-k-1}^2\right)}\right\|_d \\
&\leq \frac{1}{\beta}\sum_{k=1}^{\infty} k\left\|\left\{\frac{\beta^k\left(\omega + \alpha\epsilon_{t-k-1}^2\right)}{\omega}\right\}^{s/d}\right\|_d \\
&\leq \frac{1}{\omega^{s/d}\beta}\left\{E_{\theta_0}\left(\omega + \alpha\epsilon_1^2\right)^s\right\}^{1/d}\sum_{k=1}^{\infty} k|\beta|^{sk/d} < \infty, \qquad (9)
\end{aligned}$$

where $\|X\|_d^d = E|X|^d$ for any random variable X. The idea of exploiting the inequality $x/(1+x) \leq x^s$ for all $x > 0$ is due to Boussama (1998). Finally $\sigma_1^{-2}(\partial\sigma_1^2/\partial\theta)$ admits moments of any order, and J is well defined. The identifiability assumptions **A3** and **A4** entail the invertibility of J (see (ii) of the proof of Theorem 2.2 in Francq and Zakoïan (2004)). The consistency (7) of the QMLE, Assumption **A5** and a Taylor expansion of $\partial\tilde{\mathbf{l}}_n(\cdot)/\partial\theta$ yield

$$0 = \sqrt{n}\frac{\partial\tilde{\mathbf{l}}_n(\hat{\theta}_n^{QML})}{\partial\theta} = \sqrt{n}\frac{\partial\tilde{\mathbf{l}}_n(\theta_0)}{\partial\theta} + \left(\frac{\partial^2\tilde{\mathbf{l}}_n(\theta_{ij}^*)}{\partial\theta_i\partial\theta_j}\right)\sqrt{n}\left(\hat{\theta}_n^{QML} - \theta_0\right)$$

where the θ_{ij}^* are between $\hat{\theta}_n^{QML}$ and θ_0. The AN in (8) is then obtained by showing that

$$\begin{aligned}
\sqrt{n}\frac{\partial\tilde{\mathbf{l}}_n(\theta_0)}{\partial\theta} &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(1-\eta_t^2)\frac{1}{\sigma_t^4}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta'}(\theta_0) + o_P(1) \\
&\xrightarrow{d} \mathcal{N}\left\{0, (E\eta_1^4 - 1)J\right\}, \qquad (10)
\end{aligned}$$

and

$$n^{-1}\sum_{t=1}^{n}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\tilde{\ell}_t(\theta_{ij}^*)\to J(i,j) \text{ in probability.} \tag{11}$$

The convergence (10) follows from the central limit theorem for martingale differences given by Billingsley (1961). To show (11), a new Taylor expansion and already given arguments are employed.

It is worth-noting that no moment assumption is required for the observed process. This is particularly interesting for financial series, for which the existence of fourth and even second-order moments is questionable. The moment assumption **A6** on the iid process (η_t) is obviously necessary for the existence of the variance of the Gaussian distribution in (8). In the ARCH case we find the same asymptotic variance as for the QGLSE; see (5). Note also that the iid assumption on (η_t) can be replaced by a martingale difference assumption at the price of additional moment assumptions; see Escanciano (2007).

Tables 2 and 3 illustrate the asymptotic accuracy obtained from using the LSE and the QMLE for several ARCH(1) models with Gaussian errors and parameter $\omega_0 = 1$. When a sequence of random vectors X_n converges in law to a Gaussian distribution, we denote by $\mathrm{Var}_{as}(X_n)$ the variance of this Gaussian distribution. In view of (4), the asymptotic variance of the LSE in Table 2 is explicitly given by

$$\mathrm{Var}_{as}\{\sqrt{n}(\hat{\theta}_n^{LS}-\theta_0)\} = 2A^{-1}BA^{-1},$$

where

$$A=\begin{pmatrix}1 & E_{\theta_0}\epsilon_1^2\\ E_{\theta_0}\epsilon_1^2 & E_{\theta_0}\epsilon_1^4\end{pmatrix}, \qquad B=\begin{pmatrix}E_{\theta_0}\sigma_2^4 & E_{\theta_0}\sigma_2^4\epsilon_1^2\\ E_{\theta_0}\sigma_2^4\epsilon_1^2 & E_{\theta_0}\sigma_2^4\epsilon_1^4\end{pmatrix},$$

with

$$E_{\theta_0}\epsilon_1^2=\frac{\omega_0}{1-\alpha_{01}}, \qquad E_{\theta_0}\epsilon_1^4=3E_{\theta_0}\sigma_1^4=\frac{3\omega_0^2(1+\alpha_{01})}{(1-3\alpha_{01}^2)(1-\alpha_{01})}.$$

The other terms of the matrix B are obtained using $\sigma_2^4=(\omega_0+\alpha_{01}\epsilon_1^2)^2$ and computing the moments of order 6 and 8 of ϵ_1^2. For an ARCH(1) model, the asymptotic variance of the QMLE is given by

$$\mathrm{Var}_{as}\{\sqrt{n}(\hat{\theta}_n^{QML}-\theta_0)\}=2J^{-1}, \quad J=E_{\theta_0}\begin{pmatrix}\frac{1}{(\omega_0+\alpha_{01}\epsilon_1^2)^2} & \frac{\epsilon_1^2}{(\omega_0+\alpha_{01}\epsilon_1^2)^2}\\ \frac{\epsilon_1^2}{(\omega_0+\alpha_{01}\epsilon_1^2)^2} & \frac{\epsilon_1^4}{(\omega_0+\alpha_{01}\epsilon_1^2)^2}\end{pmatrix},$$

but it seems impossible to obtain J explicitly as a function of $\theta_0=(\omega_0,\alpha_{01})'$. For this reason, the asymptotic variance in Table 3 is approximated by $2\hat{J}^{-1}$, where

$$\hat{J}^{-1}=\frac{1}{N}\sum_{t=1}^{N}\begin{pmatrix}\frac{1}{(\omega_0+\alpha_{01}\epsilon_t^2)^2} & \frac{\epsilon_t^2}{(\omega_0+\alpha_{01}\epsilon_t^2)^2}\\ \frac{\epsilon_t^2}{(\omega_0+\alpha_{01}\epsilon_t^2)^2} & \frac{\epsilon_t^4}{(\omega_0+\alpha_{01}\epsilon_t^2)^2}\end{pmatrix},$$

and $\epsilon_1, \ldots, \epsilon_N$ is a simulation of length $N = 10,000$ of the ARCH(1) model with parameter θ_0 and the $\mathcal{N}(0,1)$ distribution for η_t. Due to the moment conditions the asymptotic variance of the LSE does not exist for $\alpha_{01} > 0.312$ (see Table 1). Even when α_{01} is sufficiently small so that all moments exist up to a sufficiently large order, the asymptotic accuracy is much better with the QMLE than with the LSE.

Table 2 Asymptotic covariance matrix of the LSE of an ARCH(1) model

α_{01}	0.1	0.2	0.3
$\mathrm{Var}_{as}\{\sqrt{n}(\hat{\theta}_n^{LS} - \theta_0)\}$	$\begin{pmatrix} 3.98 & -1.85 \\ -1.85 & 2.15 \end{pmatrix}$	$\begin{pmatrix} 8.03 & -5.26 \\ -5.26 & 5.46 \end{pmatrix}$	$\begin{pmatrix} 151.0 & -106.5 \\ -106.5 & 77.6 \end{pmatrix}$

Table 3 Approximation of the asymptotic variance of an ARCH(1) QMLE

α_{01}	0.1	0.5	0.95
$\widehat{\mathrm{Var}}_{as}\{\sqrt{n}(\hat{\theta}_n^{QML} - \theta_0)\}$	$\begin{pmatrix} 3.46 & -1.34 \\ -1.34 & 1.87 \end{pmatrix}$	$\begin{pmatrix} 4.85 & -2.15 \\ -2.15 & 3.99 \end{pmatrix}$	$\begin{pmatrix} 6.61 & -2.83 \\ -2.83 & 6.67 \end{pmatrix}$

In passing we mention that Jensen and Rahbek (2004a) considers the QMLE $\hat{\alpha}$ in ARCH(1) models of the form $h_t = \omega_0 + \alpha_{01}\epsilon_{t-1}^2$, when the scale parameter ω_0 is known. In Jensen and Rahbek (2004a), (2004b), consistency and AN of $\hat{\alpha}$ are established even when α_{01} is outside the strict stationarity region. Although the assumption that ω_0 is known does not correspond to any realistic situation, these results are interesting from a theoretical point of view.

3.2 ARMA–GARCH models

Assuming that the log-returns follow a GARCH model may be found restrictive. The autocorrelations of certain log-returns are incompatible with a GARCH model, and lead practitioners to specify the conditional mean. In this section we limit ourselves to ARMA specifications with GARCH errors. The GARCH process is not directly observed and the observations, which represent log-returns, are now denoted by $r_1, \ldots, r_n$. The (r_t) process satisfies an ARMA(P,Q)-GARCH(p,q) model of the form

$$\begin{cases} r_t - c_0 = \sum_{i=1}^{P} a_{0i}(r_{t-i} - c_0) + \epsilon_t - \sum_{j=1}^{Q} b_{0j}\epsilon_{t-j} \\ \epsilon_t = \sqrt{h_t}\eta_t \\ h_t = \omega_0 + \sum_{i=1}^{q} \alpha_{0i}\epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_{0j} h_{t-j} \end{cases} \tag{12}$$

where (η_t) and the coefficients ω_0, α_{0i} and β_{0j} are defined as in (1), and where c_0, a_{0i} and b_{0j} are real parameters. If one allows for an ARMA part, one considerably extends the range of applications, but this approach also entails serious technical difficulties in the proof of asymptotic results. References for the estimation of ARMA-GARCH processes are Francq and Zakoïan (2004), Ling and Li (1997), Ling and Li (1998), Ling and McAleer (2003).

In Francq and Zakoïan (2004) it is shown that the consistency of the QMLE holds under assumptions similar to the pure GARCH case. In particular, the observed process does not need a finite variance for the QMLE to be consistent. However the assumption $E\eta_1 = 0$ is required.

The extension of the AN is more costly in terms of moments. This is not very surprising since in the case of pure ARMA models with iid innovations, the QMLE is asymptotically normal only when these innovations admit second-order moments; see Brockwell and Davis (1991). With GARCH innovations the AN is established in Francq and Zakoïan (2004) under a fourth-moment condition on the observed process or equivalently on the GARCH process.

4 Efficient Estimation

An important issue is the possible efficiency loss of the QMLE, resulting from the use of an inappropriate Gaussian error distribution. In practice, the true error distribution is if course unknown and the MLE cannot be computed. However, it is interesting to consider the MLE in comparison with the QMLE, as a gauge of (in)efficiency. In particular we will see that, contrary to common belief, the QMLE can be efficient even if the underlying error distribution is not Gaussian.

In this section we limit ourselves to pure GARCH models. The proof of the results of this section can be found in Francq and Zakoïan (2006). See also Berkes and Horváth (2004), Straumann (2005) for results in a more general setting.

We assume that the error process (η_t) is iid, endowed with a positive density f which is known. Conditionally on initial values, the likelihood is given by

$$L_{n,f}(\theta) = L_{n,f}(\theta; \epsilon_1, \ldots, \epsilon_n) = \prod_{t=1}^{n} \frac{1}{\tilde{\sigma}_t} f\left(\frac{\epsilon_t}{\tilde{\sigma}_t}\right).$$

A MLE of θ is defined as any measurable solution $\hat{\theta}_n^{ML}$ of

$$\hat{\theta}_n^{ML} = \arg\max_{\theta \in \Theta} L_{n,f}(\theta). \tag{13}$$

Recall that f is supposed to be positive. Assume that f is derivable and write $g(y) = yf'(y)/f(y)$. The following conditions on the smoothness of f and g are introduced:

A7: There is a $\delta_1 > 0$ such that $\sup_{y\in\mathbb{R}} |y|^{1\pm\delta_1} f(y) < \infty$;

A8: There exist $0 < C_0, \delta_2 < \infty$ such that $|g(y)| \leq C_0(|y|^{\delta_2} + 1)$ for all $y \in (-\infty, \infty)$.

Such conditions are obviously satisfied for the standard normal distribution. For the Student distribution with ν degree of freedom, we have $f(x) = K(y^2 + \nu)^{-(1+\nu)/2}$ where K is a positive constant and $g(y) = -y^2(1+\nu)/(y^2+\nu)$. Assumptions **A7** and **A8** are thus satisfied with $\nu > 0$, for $0 < \delta_1 \leq \min\{\nu, 1\}$ and $\delta_2 \geq 0$. Under **A1, A2, A4, A7, A8** the ML estimator is strongly consistent,

$$\hat{\theta}_n^{ML} \to \theta_0, \qquad \text{a.s. as} \quad n \to \infty.$$

It should be noted that no moment assumption is needed for the iid process (η_t). For the QMLE, it was crucial to assume the existence of the first two moments, and an assumption such as $E\eta_1^2 = 1$ was required for identifiability reasons. Here, because the density f is fixed, there is no identification problem. For instance, the volatility $\sqrt{h_t}$ can not be multiplied by a positive constant $c \neq 1$ and the noise η_t with density f can not be changed in the new noise $\eta_t^* = \eta_t/c$, because the density of η_t^* would not be f. Obviously when the assumption $E\eta_1^2 = 1$ is relaxed, h_t is no more the conditional variance of ϵ_t given the past, but as in Berkes and Horváth (2004), one can interpret h_t as a conditional scaling parameter of ϵ_t. The assumption that the density f is entirely known is clearly not realistic for the applications. Straumann (2005) considers the situation where the density f belongs to a known class of densities parameterized by a nuisance parameter ν, for instance a normalized Student distribution St_ν with ν degrees of freedom and unit variance. Berkes and Horváth (2004) consider a very general framework in which the function f involved in the definition (13) is not necessarily the true density of η_t. Under some regularity assumptions, Straumann (2005) and Berkes and Horváth (2004) showed that this (non Gaussian) QMLE converges almost surely to

$$\theta_0^* = (d\omega, d\alpha_1, \ldots, d\alpha_q, \beta_1, \ldots, \beta_p)', \quad d > 0. \tag{14}$$

When the density f is misspecified and non Gaussian, d is generally not equal to 1 and $\hat{\theta}_n^{ML}$ is inconsistent.

For the asymptotic normality of the MLE, it is necessary to strengthen the smoothness assumptions in **A7** and **A8**. Assume that g is twice derivable and let $g^{(0)}(y) = g(y)$, $g^{(1)}(y) = g'(y)$ and $g^{(2)}(y) = g''(y)$.

A9: There is $0 < C_0 < \infty$ and $0 \leq \kappa < \infty$ such that $|y^k g^{(k)}(y)| \leq C_0(|y|^\kappa + 1)$ for all $y \in (-\infty, \infty)$ and such that $E|\eta_1|^\kappa < \infty$ for $k = 0, 1, 2$.

A10: $\tilde{I}_f = \int \{1 + g(y)\}^2 f(y)dy < \infty$, and $\lim_{y\to\pm\infty} y^2 f'(y) = 0$.

The assumptions on the density f are mild and are satisfied for various standard distributions, such as (i) the standard Gaussian distribution, for any $\delta_1 \in (0,1]$, $\delta_2 \geq 2$ and $\kappa \geq 2$; (ii) the Student distribution with parameter $\nu > 0$, for $\delta_1 \leq \min\{\nu, 1\}$, $\delta_2 \geq 0$ and $\kappa < \nu$; (iii) the density displayed in (16) below with $\delta_1 \leq 2a$, $\delta_2 \geq 2$ and $\kappa \geq 2$. If **A1, A2, A4, A5** and **A7-A10** hold, then

$$\sqrt{n}\left(\hat{\theta}_n^{ML} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{4}{\tilde{I}_f} J^{-1}\right), \qquad \text{as} \qquad n \to \infty. \tag{15}$$

It is worth-noting that, contrary to the QMLE (see Berkes and Horváth (2003)), the MLE can be $\sqrt{n}$-consistent even when $E\eta_1^4 = \infty$. The asymptotic distributions in (8) and (15) allow to quantify the efficiency loss due to the use of Gaussian likelihood. The asymptotic variances differ only by a scaling factor, which is independent of the GARCH orders and coefficients. Interestingly, the QMLE is *not always inefficient* when the error distribution is not normal. More precisely, under the assumptions required for (8) and (15), the QMLE has the same asymptotic variance as the MLE when the density of η_t is of the form

$$f(y) = \frac{a^a}{\Gamma(a)} \exp(-ay^2)|y|^{2a-1}, \quad a > 0, \quad \Gamma(a) = \int_0^\infty t^{a-1} \exp(-t)dt. \tag{16}$$

Figure 1 displays the graph of this density for different values of a. When the density f does not belong to this family of distributions, the QMLE is asymptotically inefficient in the sense that

$$\mathrm{Var}_{as}\sqrt{n}\left\{\hat{\theta}_n^{QML} - \theta_0\right\} - \mathrm{Var}_{as}\sqrt{n}\left\{\hat{\theta}_n^{ML} - \theta_0\right\} = \left(E\eta_1^4 - 1 - \frac{4}{\tilde{I}_f}\right) J^{-1}$$

is positive definite. Table 4 illustrates the loss of efficiency of the QMLE in the case of the Student distribution with ν degrees of freedom (rescaled so that they have the required unit variance). The Asymptotic Relative Efficiency (ARE) of the MLE with respect to the QMLE is (for $\nu > 3$)

$$ARE = \mathrm{Var}_{as}\sqrt{n}\left(\hat{\theta}_n^{QML} - \theta_0\right)\left\{\mathrm{Var}_{as}\sqrt{n}\left(\hat{\theta}_n^{QML} - \theta_0\right)\right\}^{-1} = \frac{\nu(\nu-1)}{\nu(\nu-1)-12}.$$

An efficient estimator can be constructed from the QMLE in two steps. The method consists, in a first step, of running one Newton-Raphson iteration with the QMLE, or any other $\sqrt{n}$-consistent preliminary estimator $\tilde{\theta}_n$ of θ_0, as starting point: $\sqrt{n}(\tilde{\theta}_n - \theta_0) = O_P(1)$. The second step does not require any optimization procedure. Let $\hat{I}_{n,f}$ be any weakly consistent estimator of $I_f(\theta_0)$. Then the sequence $(\bar{\theta}_n)$ defined by

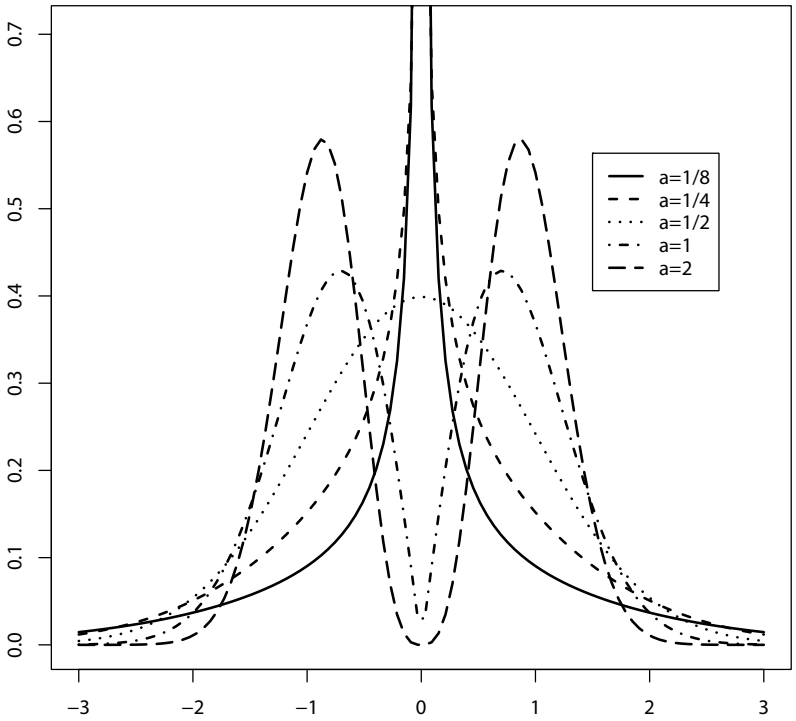

Fig. 1 Graph of the density defined by (16) for several values of $a > 0$. When η_t has a density of this form the QMLE has the same asymptotic variance as the MLE.

$$\bar{\theta}_n = \bar{\theta}_{n,f} = \tilde{\theta}_n + \hat{I}_{n,f}^{-1} \frac{1}{n} \frac{\partial}{\partial \theta} \log L_{n,f}(\tilde{\theta}_n)$$

has the same asymptotic distribution (15) as the MLE, under the same

Table 4 ARE of the MLE with respect to the QMLE when the density f of η_t is the normalized Student distribution with ν degrees of freedom and unit variance: $f(y) = \sqrt{\nu/\nu-2} f_\nu(y\sqrt{\nu/\nu-2})$, where f_ν denotes the standard Student density with ν degrees of freedom

ν	5	6	7	8	9	10	20	30	∞
ARE	2.5	1.66	1.4	1.27	1.2	1.15	1.03	1.01	1

assumptions. In concrete situations, f is unknown and $\bar{\theta}_n$ is not feasible. A feasible estimator is obtained by replacing the unknown error density f by an estimator, which can be obtained from the standardized residuals $\hat{\eta}_t = \epsilon_t / \sigma_t(\hat{\theta}_n^{QML})$, $t = 1, \ldots, n$. A non parametric kernel density estimator $\hat{f}$

can for instance be used. An issue is whether $\bar{\theta}_{n,\hat{f}}$ is an *adaptive* estimator, in the sense that it inherits the asymptotic optimality properties of $\bar{\theta}_{n,f}$. Adaptive estimation in GARCH models has been studied by several authors; see *e.g* Drost and Klaassen (1997), Engle and González-Rivera (1991), Ling and McAleer (2003), Linton (1993). From these references, adaptiveness holds in the sense that the volatility parameters can be estimated *up to a scale parameter*, with the same asymptotic precision as if the error distribution were known; see Drost and Klaassen (1997). However, adaptive estimation of all GARCH coefficients is not possible. Efficiency losses of the QMLE and semi-parametric estimators, with respect to the MLE, are quantified in González-Rivera and Drost (1999) and illustrated numerically in Drost and Klaassen (1997), Engle and González-Rivera (1991).

5 Alternative Estimators

It is known that parameter estimation is not standard for ARMA models with infinite variance innovations; see Mikosch et al. (1995). Indeed, with the notation of Section 3.2, the score vector $\epsilon_t \partial \epsilon_t / \partial \vartheta$ has a finite variance I when $E\epsilon_1^2 < \infty$ and the ϵ_t are iid. In the presence of conditionally heteroscedastic innovations, or more generally when the ϵ_t are not iid, the existence of fourth-order moments is required for the existence of I. Thus the moment condition $Er_1^4 < \infty$ seems necessary for the asymptotic normality of the LSE of the ARMA-GARCH models defined by (12). Similarly, it can be shown that the variance of the quasi-maximum likelihood score vector may not exist when $Er_1^2 = +\infty$. We have seen in Section 3.2 that moment conditions are not needed for the consistency of the QMLE. For statistical inference, consistency is however not sufficient, and the asymptotic distribution of the estimator is generally required. The asymptotic distributions of the LSE and QMLE are unknown when $Er_1^4 = +\infty$. Sections 5.1 and 5.2 present alternative estimators which require less moment assumptions on the observed process r_t. The estimators defined in Sections 5.3 and 5.4 allow one to reduce the moment assumptions on the iid process (η_t). Section 5.5 is devoted to the Whittle estimator. It will be seen that this estimator is less attractive for GARCH models than for ARMA models. The moment estimators mentioned in Section 5.6 seem particularly interesting to allow for GARCH-type effects without imposing a fully specified model. To save space we only present the main ideas of these estimation methods. The precise assumptions and asymptotic variance matrices can be found in the corresponding references.

5.1 Self–weighted LSE for the ARMA parameters

To estimate the ARMA parameters

$$\vartheta_0 = (c_0, a_{01}, \dots a_{0P}, b_{01}, \dots, b_{0Q})'$$

of the ARMA-GARCH model (12), Ling (2003) considered the self-weighted LSE (SWL) defined by

$$\hat{\vartheta}_n^{SWL} = \arg\min_{\vartheta \in \Psi} n^{-1} \sum_{t=1}^{n} \omega_t^2 \tilde{\epsilon}_t^2(\vartheta),$$

where the weights ω_t are positive measurable functions of $r_{t-1}, r_{t-2}, \dots$, Ψ is a compact subspace of $\mathbb{R}^{P+Q+1}$, and $\tilde{\epsilon}_t(\vartheta)$ are the ARMA residuals computed for the value ϑ of the ARMA parameter and with fixed initial values. Take for instance $\omega_t^{-1} = 1 + \sum_{k=1}^{t-1} k^{-1-1/s} |r_{t-k}|$ with $E|r_1|^{2s} < \infty$ and $s \in (0,1)$. It can be shown that there exist constants $K > 0$ and $\rho \in (0,1)$ such that

$$|\tilde{\epsilon}_t| \leq K\,(1 + |\eta_t|) \left(1 + \sum_{k=1}^{t-1} \rho^k |r_{t-k}|\right) \quad \text{and} \quad \left|\frac{\partial \tilde{\epsilon}_t}{\partial \vartheta_i}\right| \leq K \sum_{k=1}^{t-1} \rho^k |r_{t-k}|.$$

It follows that

$$|\omega_t \tilde{\epsilon}_t| \leq K\,(1 + |\eta_t|) \left(1 + \sum_{k=1}^{\infty} k^{1+1/s} \rho^k\right), \; \left|\omega_t \frac{\partial \tilde{\epsilon}_t}{\partial \vartheta_i}\right| \leq K \left(1 + \sum_{k=1}^{\infty} k^{1+1/s} \rho^k\right).$$

Thus

$$E \left|\omega_t^2 \tilde{\epsilon}_t \frac{\partial \tilde{\epsilon}_t}{\partial \vartheta_i}\right|^2 \leq K^4 E\,(1 + |\eta_1|)^2 \left(\sum_{k=1}^{\infty} k^{1+1/s} \rho^k\right)^4 < \infty,$$

which entails a finite variance for the SWL score vector $\omega_t^2 \tilde{\epsilon}_t \partial \tilde{\epsilon}_t / \partial \vartheta$. Ling (2006) then deduced the asymptotic normality of $\sqrt{n}(\hat{\vartheta}_n^{SWL} - \vartheta_0)$, allowing for the case $Er_1^2 = \infty$.

5.2 Self–weighted QMLE

To obtain an AN estimator of the parameter $\varphi_0 = (\vartheta_0', \theta_0')'$ in the ARMA-GARCH model (12) under mild moment assumptions on the observed process, Ling (2006) proposed the self-weighted QMLE

$$\hat{\varphi}_n^{SWQ} = \arg\min_{\varphi \in \Phi} n^{-1} \sum_{t=1}^{n} \omega_t \tilde{\ell}_t(\varphi),$$

where $\tilde{\ell}_t(\varphi) = \tilde{\epsilon}_t^2(\vartheta)/\tilde{\sigma}_t^2(\varphi) + \log \tilde{\sigma}_t^2(\varphi)$ with obvious notations. To understand the principle of this estimator, let us note that the minimized criterion generally converges to a limit criterion $\mathbf{l}(\varphi) = E_\varphi \omega_t \ell_t(\varphi)$ satisfying

$$\mathbf{l}(\varphi) - \mathbf{l}(\varphi_0) = E_{\varphi_0}\omega_t \left\{ \log \frac{\sigma_t^2(\varphi)}{\sigma_t^2(\varphi_0)} + \frac{\sigma_t^2(\varphi_0)}{\sigma_t^2(\varphi)} - 1 \right\} + E_{\varphi_0}\omega_t \frac{\{\epsilon_t(\vartheta) - \epsilon_t(\vartheta_0)\}^2}{\sigma_t^2(\varphi)} + E_{\varphi_0}\omega_t \frac{2\eta_t \sigma_t(\varphi_0)\{\epsilon_t(\vartheta) - \epsilon_t(\vartheta_0)\}}{\sigma_t^2(\varphi)}.$$

The last expectation (when it exists) is zero because η_t is centered and is independent of the other random variables involved in the expectation. From the inequality $x - 1 \geq \log x$, we have

$$E_{\varphi_0}\omega_t \left\{ \log \frac{\sigma_t^2(\varphi)}{\sigma_t^2(\varphi_0)} + \frac{\sigma_t^2(\varphi_0)}{\sigma_t^2(\varphi)} - 1 \right\} \geq E_{\varphi_0}\omega_t \left\{ \log \frac{\sigma_t^2(\varphi)}{\sigma_t^2(\varphi_0)} + \log \frac{\sigma_t^2(\varphi_0)}{\sigma_t^2(\varphi)} \right\}.$$

Hence under the usual identifiability assumptions, $\mathbf{l}(\varphi) \geq \mathbf{l}(\varphi_0)$ with equality if and only if $\varphi = \varphi_0$. Note that the orthogonality between η_t and the weight ω_t is essential.

Ling (2006) showed consistency and AN of $\hat{\varphi}_n^{SWQ}$ under the assumption $E|r_1|^s < \infty$ for some $s > 0$.

5.3 L_p–estimators

The weighted estimators of the previous sections require the moment assumption $E\eta_1^4 < \infty$. Practitioners often claim that financial series do not admit (even low-order) moments. In GARCH processes an infinite variance can be obtained either by relaxing the parameters constraint or by allowing an infinite variance for η_t. In the GARCH(1,1) case the two sets of assumptions

$$i) : \begin{cases} \alpha_{01} + \beta_{01} \geq 1 \\ E\eta_1^2 = 1 \end{cases} \quad \text{or} \quad ii) : E\eta_1^2 = \infty$$

imply an infinite variance for ϵ_t. Under i), and the strict stationarity assumption, the asymptotic distribution of the QLME is generally Gaussian (see Section 3), whereas the usual estimators have non standard asymptotic distributions or are even non-consistent under ii); see Berkes and Horváth (2003), Hall and Yao (2003), Mikosch and Straumann (2002). It is therefore of interest to define alternative estimators enjoying a Gaussian asymptotic distribution under ii), or even under the more general situation where both $\alpha_{01} + \beta_{01} > 1$ and $E\eta_1^2 = \infty$ are allowed for.

Note that a GARCH model is generally defined under the standardization $E\eta_1^2 = 1$. When the existence of $E\eta_1^2$ is relaxed, one can identify the GARCH coefficients by imposing that the median of η_1^2 be $\tau = 1$. In the framework of

ARCH(q) models, Horváth and Liese (2004) consider L_p-estimators, including the L_1-estimator

$$\hat{\theta}_n^{L_1} = \arg\min_{\theta} n^{-1} \sum_{t=1}^{n} \omega_t \left| \epsilon_t^2 - \omega - \sum_{i=1}^{q} \alpha_i \epsilon_{t-1}^2 \right|,$$

where, for example, $\omega_t^{-1} = 1 + \sum_{i=1}^{p} \epsilon_{t-i}^2 + \epsilon_{t-i}^4$. When η_t^2 has a density, continuous and positive around its median $\tau = 1$, the consistency and asymptotic normality of $\hat{\theta}_n^{L_1}$ is shown in Horváth and Liese (2004), without any moment assumption.

5.4 Least absolute deviations estimators

In the framework of ARCH and GARCH models, Peng and Yao (2003) studied several least absolute deviations estimators. An interesting specification is the following

$$\hat{\theta}_n^{LAD} = \arg\min_{\theta} n^{-1} \sum_{t=1}^{n} \left| \log \epsilon_t^2 - \log \tilde{\sigma}_t^2(\theta) \right|. \tag{17}$$

With this estimator it is convenient to define the GARCH parameters under the condition that the median of η_1^2 is equal to 1. It entails a reparametrization of standard GARCH models. Consider, for instance, a GARCH(1,1) model with parameters ω_0, α_{01} and β_{01}, and a Gaussian noise η_t. Since the median of η_1^2 is $\tau = 0.4549...$, the median of the square of $\eta_t^* = \eta_t/\sqrt{\tau}$ is 1, and the model is rewritten as

$$\epsilon_t = \sigma_t \eta_t^*, \quad \sigma_t^2 = \tau\omega_0 + \tau\alpha_{01}\epsilon_{t-1}^2 + \beta_{01}\sigma_{t-1}^2.$$

It is interesting to note that the error terms $\log \eta_t^{*2} = \log \epsilon_t^2 - \log \tilde{\sigma}_t^2(\theta)$ are iid with median 0 when $\theta = \theta_0$. Intuitively, this is the reason why it is not necessary to use weights in the sum (17). Under the moment assumption $E\epsilon_1^2 < \infty$ and certain regularity assumptions, it is shown in Peng and Yao (2003) that there exists a local solution of (17) which is weakly consistent and AN, with the standard rate of convergence $n^{1/2}$. This convergence holds even in the case of heavy-tailed errors : no condition on the moments of η_1 beyond $E\eta_1^2 = 1$ is imposed.

5.5 Whittle estimator

Whittle estimation is a standard method for ARMA models, working in the spectral domain of the process; see Brockwell and Davis (1991), Section 10.8 for further details. It is well known that, under the moment assumption $E\epsilon_1^4 < \infty$, the square of a GARCH(p,q) model satisfies an ARMA$(p \wedge q, q)$ model

$$\phi_{\theta_0}(L)\epsilon_t^2 = \omega_0 + \psi_{\theta_0}(L)u_t, \tag{18}$$

where L denotes the lag operator,

$$\phi_{\theta_0}(z) = 1 - \sum_{i=1}^{p\wedge q}(\alpha_{0i} + \beta_{0i})z^i, \quad \psi_{\theta_0}(z) = 1 - \sum_{i=1}^{p}\beta_{0i}z^i, \quad u_t = (\eta_t^2 - 1)\sigma_t^2.$$

Thus, the spectral density of ϵ_t^2 is

$$f_{\theta_0}(\lambda) = \frac{Eu_t^2}{2\pi}\frac{\left|\psi_{\theta_0}(e^{-i\lambda})\right|^2}{\left|\phi_{\theta_0}(e^{-i\lambda})\right|^2}.$$

Denote by $\hat{\gamma}_{\epsilon^2}(h)$ the sample autocovariance of ϵ_t^2 at lag h. At the Fourier frequencies $\lambda_j = 2\pi j/n \in (-\pi, \pi]$, the periodogram

$$I_n(\lambda_j) = \sum_{|h|<n}\hat{\gamma}_{\epsilon^2}(h)e^{-ih\lambda_j}, \quad j \in \mathfrak{J} = \left\{\left[-\frac{n}{2}\right] + 1, \ldots, \left[\frac{n}{2}\right]\right\},$$

can be considered as a non parametric estimator of $2\pi f_{\theta_0}(\lambda_j)$. Let

$$u_t(\theta) = \frac{\phi_\theta(L)}{\psi_\theta(L)}\left\{\epsilon_t^2 - \omega\phi_\theta^{-1}(1)\right\}.$$

It can be shown that

$$Eu_1^2(\theta) = \frac{Eu_1^2(\theta_0)}{2\pi}\int_{-\pi}^{\pi}\frac{f_{\theta_0}(\lambda)}{f_\theta(\lambda)}d\lambda \geq Eu_1^2(\theta_0)$$

with equality if and only if $\theta = \theta_0$; see Brockwell and Davis (1991) Proposition 10.8.1. In view of this inequality, it seems natural to consider the so-called Whittle estimator

$$\hat{\theta}_n^W = \arg\min_\theta \frac{1}{n}\sum_{j\in\mathfrak{J}}\frac{I_n(\lambda_j)}{f_\theta(\lambda_j)}.$$

For ARMA models with iid innovations the Whittle estimator has the same asymptotic behavior as the QMLE and LSE. For GARCH processes the Whittle estimator has still the same asymptotic behavior as the LSE, but simulations studies indicate that the Whittle estimator, for normal and student noises (η_t), is less accurate than the QMLE. Moreover Giraitis and Robinson

(2001), Mikosch and Straumann (2002), Straumann (2005) have shown that consistency requires the existence of $E\epsilon_1^4$, and asymptotic normality requires $E\epsilon_1^8 < \infty$.

5.6 Moment estimators

A sequence (ϵ_t) is called *weak* white noise if the ϵ_t's are centered and uncorrelated, but not necessarily independent. In contrast, a sequence of centered and independent random variables is sometimes called *strong* white noise. The GARCH process is a leading example of weak white noise, but there exist numerous other examples of weak white noises satisfying (18). Consider for example the process $v_t = \eta_t\eta_{t-1}$ where (η_t) is iid $\mathcal{N}(0,1)$. This process is clearly weak white noise. Straightforward computations show that v_t^2 satisfies a weak MA(1) representation of the form $v_t^2 = 1 + u_t + \theta u_{t-1}$, where (u_t) is weak white noise. Although (v_t) does not belong to the class of the *strong* GARCH models defined by (1), it can be called *weak* GARCH, in the sense that (v_t) is a white noise and (v_t^2) satisfies an ARMA model. The ARMA representations (18) of these weak GARCH models are estimated in Francq and Zakoïan (2000) by LS, under moment and mixing conditions, but without imposing a particular parametric model for (ϵ_t).

The generalized method of moment (GMM) approach is particularly relevant (see Rich et al. (1991)) to estimate ARCH models without assuming strong assumptions on the noise (η_t).

To finish this non exhaustive list of alternative GARCH estimators, let us mention the existence of Bayesian estimators, using Monte Carlo integration with importance sampling for the computation of the posterior expectations; see Geweke (1989).

6 Properties of Estimators when some GARCH Coefficients are Equal to Zero

To obtain the AN of the QMLE of GARCH models, a crucial assumption is that the true parameter vector has strictly positive components. When some components are equal to zero, the parameter, which is constrained to have nonnegative components, lies at the boundary of the parameter space and then, Assumption **A5** in Section 3.1 is not satisfied. This assumption is a serious limitation to the estimation theory of GARCH. Indeed it could be particularly useful to derive the asymptotic distribution of the QMLE of a GARCH(p,q) model when, for instance, the underlying process is a GARCH$(p-1,q)$, or a GARCH$(p,q-1)$ process. Tests of the significance of the coefficients and tests of conditional homoscedasticity constitute typical

situations where we have to study the QMLE when the parameter is at the boundary.

In this section we study the asymptotic behaviour of the QMLE for GARCH processes, when the true parameter may have zero coefficients. We first see, by means of an elementary example, why the asymptotic distribution of the QMLE cannot be Gaussian when one or several GARCH coefficients are equal to zero.

6.1 Fitting an ARCH(1) model to a white noise

The QMLE of an ARCH(1) model is obtained by minimizing the criterion

$$\mathbf{l}_n(\omega,\alpha) = n^{-1}\sum_{t=2}^{n} \ell_t(\omega,\alpha), \qquad \ell_t(\omega,\alpha) = \frac{\epsilon_t^2}{\sigma_t^2} + \log\sigma_t^2,$$

where $\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2$. In absence of constraints on the coefficients, the value of σ_t^2 could be negative (this is the case when $\alpha < 0$, $\omega > 0$ and $\epsilon_{t-1}^2 > -\omega/\alpha$). In such a situation, $\ell_t(\omega,\alpha)$, and thus the objective function $\mathbf{l}_n(\omega,\alpha)$, are not defined. This is the reason why the minimization is made under the constraints $\omega > 0$ and $\alpha \geq 0$. The QMLE estimator

$$(\hat{\omega}_n, \hat{\alpha}_n) = \underset{\omega>0,\, \alpha\geq 0}{\arg\min}\, \mathbf{l}_n(\omega,\alpha)$$

then satisfies $\hat{\alpha}_n \geq 0$ almost surely, for all n. When the process is a white noise, then $\alpha_{01} = 0$ and with probability one

$$\sqrt{n}(\hat{\alpha}_n - \alpha_{01}) = \sqrt{n}\hat{\alpha}_n \geq 0, \qquad \forall n.$$

In this case $\sqrt{n}(\hat{\alpha}_n - \alpha_{01})$ cannot converge in law to any non-degenerate Gaussian distribution $\mathcal{N}(m, s^2)$ with $s^2 > 0$. Indeed

$$\lim_{n\to\infty} P\left\{\sqrt{n}(\hat{\alpha}_n - \alpha_{01}) < 0\right\} = 0 \quad \text{whereas} \quad P\left\{\mathcal{N}(m,s^2) < 0\right\} > 0.$$

For the same reason, when the true value of a general GARCH parameter has zero components, the asymptotic distribution cannot be Gaussian, for the QMLE or for any other estimator which takes into account the positivity constraints.

6.2 On the need of additional assumptions

To prove the existence of the information matrix involved in the asymptotic distribution of the QMLE, we have to show that the variance of the vector $\sigma_t^{-2}(\theta_0)\partial\sigma_t^2(\theta_0)/\,\partial\theta$, and the expectation of the matrix

$$J_t = \frac{1}{\sigma_t^4(\theta_0)}\left(\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta'}\right)$$

are finite. A bound for these norms can be shown to be of the form Kc^{-1} or Kc^{-2}, where K is a constant and $c>0$ is the smallest component of θ_0. Obviously, the proof breaks down when one or several components of θ_0 are equal to zero.

To see this technical problem more clearly, let us consider the ARCH(1) example. If $\omega_0\alpha_{01}>0$ then the expectation of J_t is finite because

$$EJ_t = E\frac{1}{(\omega_0+\alpha_{01}\epsilon_1^2)^2}\begin{pmatrix}1 & \epsilon_1^2\\ \epsilon_1^2 & \epsilon_1^4\end{pmatrix} \le \begin{pmatrix}\omega_0^{-2} & \omega_0^{-1}\alpha_{01}^{-1}\\ \omega_0^{-1}\alpha_{01}^{-1} & \alpha_{01}^{-2}\end{pmatrix},$$

where the last inequality has to be taken componentwise. However, if $\alpha_{01}=0$

$$EJ_t = \frac{1}{\omega_0^2}E\begin{pmatrix}1 & \epsilon_1^2\\ \epsilon_1^2 & \epsilon_1^4\end{pmatrix}$$

is finite when $E\epsilon_1^4<\infty$ only.

Such extra moment assumptions seem necessary for ARCH models and for the GARCH(1,1), but can sometimes be avoided for more complex GARCH models. Consider for example a strictly stationary GARCH(p,q) process with $\alpha_{01}>0$ and $\beta_{01}>0$. Then, because $\sum_{j=1}^p\beta_{0j}<1$, the following ARCH(∞) expansion holds $\sigma_t^2(\theta_0)=c_0+\sum_{j=1}^\infty b_{0j}\epsilon_{t-j}^2$ with $c_0>0$ and $b_{0j}>0$ for all j; see Giraitis et al. (2008) for a review on ARCH(∞) models. Similar expansions hold for the derivatives $\partial\sigma_t^2/\partial\theta_i$. Thus every term ϵ_{t-j}^2 appearing in the numerator of this ratio $\{\partial\sigma_t^2/\partial\theta\}/\sigma_t^2$ is also present in the denominator. In such a situation the moment assumption $E\epsilon_1^4<\infty$ is not necessary for the existence of EJ_t.

6.3 Asymptotic distribution of the QMLE on the boundary

For simplicity, let us take a parameter space of the form

$$\Theta = [\underline{\omega},\overline{\omega}]\times[0,\overline{\alpha}_1]\times\cdots\times[0,\overline{\beta}_p]$$

where $\underline{\omega} > 0$ and $\overline{\alpha}_1, \ldots, \overline{\beta}_p > 0$. We assume that

A11: $\theta_0 \in (\underline{\omega}, \overline{\omega}) \times [0, \overline{\alpha}_1) \times \cdots \times [0, \overline{\beta}_p)$,

allowing for zero GARCH coefficients, but excluding the case where θ_0 is on the upper boundary of Θ. When Θ is not a product of intervals, Assumption **A11** must be modified appropriately. We now define the "local" parameter space

$$\Lambda = \Lambda(\theta_0) = \Lambda_1 \times \cdots \times \Lambda_{p+q+1},$$

where $\Lambda_1 = \mathbb{R}$, and, for $i = 2, \ldots, p+q+1$, $\Lambda_i = \mathbb{R}$ if $\theta_{0i} \neq 0$ and $\Lambda_i = [0, \infty)$ if $\theta_{0i} = 0$. In view of the positivity constraints, the random vector $\sqrt{n}(\hat{\theta}_n - \theta)$ belongs to Λ with probability one.

We already know that the QMLE is consistent under **A1**–**A4**, even when θ_0 is on the boundary of Θ. If in addition **A6**, **A11** and either

A12: $E\epsilon_1^6 < \infty$

or alternatively

A12': $\sigma_t^2(\theta_0) = c_0 + \sum_{j=1}^{\infty} b_{0j}\epsilon_{t-j}^2$ with $b_{0j} > 0$ for all $j \geq 1$

hold, then

$$\sqrt{n}(\hat{\theta}_n^{QML} - \theta_0) \xrightarrow{d} \lambda^{\Lambda} := \arg \inf_{\lambda \in \Lambda} \{\lambda - Z\}' J \{\lambda - Z\}, \tag{19}$$

with $Z \sim \mathcal{N}\left(0, (E\eta_1^4 - 1)J^{-1}\right)$.

When $\theta_0 \in \overset{\circ}{\Theta}$, we have $\Lambda = \mathbb{R}^{p+q+1}$ and we retrieve the standard result because $\lambda^{\Lambda} = Z \sim \mathcal{N}\left(0, (E\eta_1^4 - 1)J^{-1}\right)$. When θ_0 is on the boundary, the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n^{QML} - \theta_0)$ is more complex than a Gaussian. This is the law of the projection of the Gaussian vector Z on the convex cone Λ. The reader is referred to Andrews (1999) for similar results on a general framework, and to Francq and Zakoïan (2007) for the proof of (19). For fitting ARCH(q) models, Jordan (2003) allows a parameter belonging to the boundary of a non compact set, and a DGP which is not necessarily an ARCH process, but requires in particular the moment assumption $E\epsilon_1^8 < \infty$.

6.4 Application to hypothesis testing

An important consequence of the non Gaussian behavior of the QMLE is that the Wald and Likelihood-Ratio (LR) tests do not have the standard χ^2 asymptotic distribution. As an illustration consider the ARCH(2) case with $\theta_0 = (\omega_0, 0, 0)$. We have

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} \sim \mathcal{N}\left\{0, (E\eta_1^4 - 1)J^{-1} = \begin{pmatrix} (E\eta_1^4 + 1)\omega_0^2 & -\omega_0 & -\omega_0 \\ -\omega_0 & 1 & 0 \\ -\omega_0 & 0 & 1 \end{pmatrix}\right\}$$

and we can show that

$$\lambda^{\Lambda} = \begin{pmatrix} Z_1 + \omega Z_2^- + \omega Z_3^- \\ Z_2^+ \\ Z_3^+ \end{pmatrix}, \quad Z_i^+ = \max\{Z_i, 0\} \text{ and } Z_i^- = \min\{Z_i, 0\}.$$

We can see that, asymptotically, we have $\hat{\alpha}_1 = 0$ (or $\hat{\alpha}_2 = 0$) with probability $1/2$, and $\hat{\alpha}_1 = \hat{\alpha}_2 = 0$ with probability $1/4$. Consequently, for the test of the null hypothesis $H_0 : \alpha_1 = \alpha_2 = 0$, the Wald statistic $\mathrm{W}_n = n\left(\hat{\alpha}_1^2 + \hat{\alpha}_2^2\right)$ has a discrete component, and thus cannot be the usual χ_2^2. More precisely, it is easy to see that under H_0

$$\mathrm{W}_n \xrightarrow{d} \mathrm{W} \sim \frac{1}{4}\delta_0 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2.$$

One can show that the LR test has the same nonstandard asymptotic distribution (in the Gaussian case), whereas the Lagrange Multiplier (LM) test conserves its usual χ^2 asymptotic distribution, even when H_0 puts θ_0 on the boundary of the parameter space; see Demos and Sentana (1998), Francq and Zakoïan (2007). This is not very surprising, since the likelihood of the constrained model is equal to that of the unconstrained model when $\hat{\alpha}_1 = \hat{\alpha}_2 = 0$, but the score is not necessarily zero when $\hat{\alpha}_1 = \hat{\alpha}_2 = 0$ (see Figure 2).

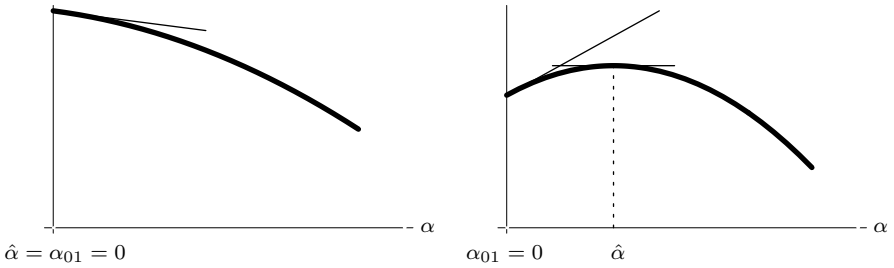

Fig. 2 Projected Log-likelihood (full line) $\alpha \mapsto \log L_n(\hat{\omega}, \alpha)$ of an ARCH(1) model with $\alpha_{01} = 0$. In the right-hand graph we have $\hat{a} > 0$, $\partial \log L_n(\hat{\omega}, \alpha)/\partial\alpha = 0$ and the score $\partial \log L_n(\hat{\omega}, 0)/\partial\alpha > 0$. In the left-hand graph we have $\hat{\alpha} = 0$ and $\partial \log L_n(\hat{\omega}, \alpha)/\partial\alpha = \partial \log L_n(\hat{\omega}, 0)/\partial\alpha < 0$. In both cases the score is almost surely non null.

Another important consequence of the non standard asymptotic distribution (19), is that the Wald, LM and LR tests do not have the same local asymptotic power. The Wald test generally outperforms the LM test in terms of local asymptotic power. This is not surprising because the LM test do not take into account the one-sided nature of the alternatives. It is of course possible to derive one-sided versions of the LM test; see *e.g.* Demos and Sentana (1998).

7 Conclusion

Since many financial series exhibit heavy-tailed marginal distributions, it is particularly important to obtain estimation procedures which do not hinge on high-order moment assumptions. The QMLE is the most popular method for estimating GARCH models. In the general ARMA-GARCH case, the consistency is obtained without moment assumption on the observed process ϵ_t, even when the parameter is on the boundary of the parameter space (a situation frequently encountered in test problems). In the pure GARCH case with $\theta_0 \in \overset{\circ}{\Theta}$ the AN is also obtained without moment assumption on ϵ_t, but addition assumptions are required in the general ARMA-GARCH case. When θ_0 is on the boundary of the parameter space, the asymptotic distribution of the QMLE is no longer Gaussian, but is that of the projection of a Gaussian vector on a convex cone. The main drawbacks of the QMLE are that i) the estimator is not explicit and it requires a numerical optimization, ii) the AN requires the existence of fourth-order moments for the iid process η_t, iii) the estimator is in general inefficient, iv) the AN requires moments assumptions on ϵ_t in the general ARMA-GARCH case, v) a fully parametric specification is required. Concerning the point iii) it is however interesting to note that the QMLE is not only efficient in the Gaussian case, but also when the distribution of η_t belongs to the class defined in Section 4. At least in the ARCH case, a two-step LSE should respond satisfactorily to the point i), but with a cost in terms of moment conditions. Weighted L_p and least absolute deviations estimators have been recently developed to alleviate the point ii). The MLE is a fully satisfactory response to the points ii) and iii), but requires a complete specification of the error distribution, unless adaptive estimators be employed. Also very recently, self-weighted LSE and self-weighted QMLE have been developed to respond to the point iv). Methods based on orthogonality conditions, such as the GMM, are simple and obviously more robust to model misspecifications, and are therefore worthwhile procedures for considering the points i) and v).

References

Andrews, D.W.K. (1999): Estimation when a parameter is on a boundary. *Econometrica* **67**,1341–1384.

Berkes, I. and Horváth, L. (2003): The rate of consistency of the quasi-maximum likelihood estimator. *Statistics and Probability Letters* **61**, 133–143.

Berkes, I. and Horváth, L. (2004): The efficiency of the estimators of the parameters in GARCH processes *Annals of Statistics* **32**, 633–655.

Berkes, I., Horváth, L. and Kokoszka, P.S. (2003): GARCH processes: structure and estimation. *Bernoulli* **9**, 201–227.

Billingsley, P. (1961): The Lindeberg-Levy theorem for martingales. *Proceedings of the American Mathematical Society* **12**, 788–792.

Bose, A. and Mukherjee, K. (2003): Estimating the ARCH parameters by solving linear equations. *Journal of Time Series Analysis* **24**, 127–136.

Bougerol, P. and Picard, N. (1992): Stationarity of GARCH processes and of some non-negative time series. *Journal of Econometrics* **52**, 115–127.

Boussama, F. (1998): *Ergodicité, mélange et estimation dans les modèles GARCH*. PhD Thesis, Université Paris-7, Paris.

Brockwell, P.J. and Davis, R.A. (1991): Time series: theory and methods. Springer, New-York.

Chen, M. and An, H.Z. (1998): A note on the stationarity and the existence of moments of the GARCH model. *Statistica Sinica* **8**, 505–510.

Demos, A. and Sentana, E. (1998): Testing for GARCH effects: a one-sided approach. *Journal of Econometrics* **86**, 97–127.

Drost, F.C. and Klaassen, C.A.J. (1997): Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* **81**, 193–221.

Engle, R.F. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica* **50**, 987–1007.

Engle, R.F. and González-Rivera, G. (1991): Semiparametric ARCH models. *Journal of Business and Econometric Statistics* **9**, 345–359.

Escanciano, J.C. (2007): Quasi-maximum likelihood estimation of semi-strong GARCH models. *Working documents*.

Francq, C. and Zakoïan, J.M. (2000): Estimating weak GARCH representations. *Econometric Theory* **16**, 692–728.

Francq, C. and Zakoïan, J.M. (2004): Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* **10**, 605–637.

Francq, C. and Zakoïan, J.M. (2006): On efficient inference in GARCH processes. *In: Bertail, P., Doukhan, P., Soulier, P. (Eds): Statistics for dependent data*, 305–327. Springer, New-York.

Francq, C. and Zakoïan, J.M. (2007): Quasi-maximum likelihood estimation in GARCH processes when some coefficients are equal to zero. *Stochastic Processes and their Applications* **117**, 1265–1284.

Giraitis, L., Leipus, R. and Surgailis, D. (2008): ARCH(∞) models and long-memory properties. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 70–84 Springer, New York.

González-Rivera, G. and Drost, F.C. (1999): Efficiency comparisons of maximum-likelihood based estimators in GARCH models. *Journal of Econometrics* **93**, 93–111.

Gouriéroux, C. (1997): *ARCH models and financial applications*. Springer, New York.

Geweke, J. (1989): Exact predictive densities for linear models with ARCH disturbances. *Journal of Econometrics* **40**, 63–86.

Giraitis, L. and Robinson, P.M. (2001): Whittle estimation of ARCH models. *Econometric Theory* **17**, 608–631.

Hall, P. and Yao, Q. (2003): Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71**, 285–317.

Hamilton, J.D. (1994): Time series analysis. *Princeton University Press, Princeton*.

Horváth, L. and Liese, F. (2004): L_p-estimators in ARCH models. *Journal of Statistical Planning and Inference* **119**, 277–309.

Jensen, S.T. and Rahbek, A. (2004a): Asymptotic normality of the QMLE estimator of ARCH in the nonstationary case. *Econometrica* **72**, 641–646.

Jensen, S.T. and Rahbek, A. (2004b): Asymptotic inference for nonstationary GARCH. *Econometric Theory* **20**, 1203–1226.

Jordan, H. (2003): *Asymptotic properties of ARCH(p) quasi maximum likelihood estimators under weak conditions*. PhD Thesis, University of Vienna.

Lee, S.W. and Hansen, B.E. (1994): Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory* **10**, 29–52.

Lindner, A. (2008): Stationarity, mixing, distributional properties and moments of GARCH(p,q)–processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 43–69. Springer, New York.

Ling, S. (2003): Self-weighted LSE and MLE for ARMA-GARCH models. Unpublished working paper, HKUST.

Ling, S. (2006): Self-weighted and local quasi-maximum likelihood estimators for ARMA-GARCH/IGARCH models. *Journal of Econometrics*. To appear.

Ling, S. and Li, W.K. (1997): On fractionally integreted autoregressive moving-average time series models with conditional heteroscedasticity. *Journal of the American Statistical Association* **92**, 1184–1194.

Ling, S. and Li, W.K. (1998): Limiting distributions of maximum likelihood estimators for unstable ARMA models with GARCH errors. *Annals of Statistics* **26**, 84–125.

Ling, S. and McAleer, M. (2003): Asymptotic theory for a vector ARMA-GARCH model. *Econometric Theory* **19**, 280–310.

Ling, S. and McAleer, M. (2003): Adaptive estimation in nonstationry ARMA models with GARCH noises. *Annals of Statistics* **31**, 642–674.

Linton, O. (1993): Adaptive estimation in ARCH models. *Econometric Theory* **9**, 539–564.

Lumsdaine, R.L. (1996): Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica* **64**, 575–596.

Mikosch, T., Gadrich, T., Klüppelberg, C. and Adler, R.J. (1995): Parameter estimation for ARMA models with infinite variance innovations. *Annals of Statistics* **23**, 305–326.

Mikosch, T. and Straumann, D. (2002): Whittle estimation in a heavy-tailed GARCH(1,1) model. *Stochastic Processes and their Application* **100**, 187–222.

Nelson, D.B. (1990): Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory* **6**, 318–334.

Peng, L. and Yao, Q. (2003): Least absolute deviations estimation for ARCH and GARCH models. *Biometrika* **90**, 967–975.

Rich, R.W., Raymond, J. and Butler, J.S. (1991): Generalized instrumental variables estimation of autoregressive conditional heteroskedastic models. *Economics Letter* **35**, 179–185.

Straumann, D. (2005): *Estimation in conditionally heteroscedastic time series models. Lecture Notes in Statistics*. Springer, Berlin Heidelberg.

Straumann, D. and Mikosch, T. (2006): Quasi-MLE in heteroscedastic time series: a stochastic recurrence equations approach. *Annals of Statistics* **34**, 2449–2495.

Wald, A. (1949): Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics* **20**, 595–601.

Weiss, A.A. (1986): Asymptotic theory for ARCH models: estimation and testing. *Econometric Theory* **2**, 107–131.

Practical Issues in the Analysis of Univariate GARCH Models

Eric Zivot

Abstract This chapter gives a tour through the empirical analysis of univariate GARCH models for financial time series with stops along the way to discuss various practical issues associated with model specification, estimation, diagnostic evaluation and forecasting.

1 Introduction

There are many very good surveys covering the mathematical and statistical properties of GARCH models. See, for example, Bera and Higgins (1995), Bollerslev et al. (1994), Pagan (1996), Palm (1996), Diebold and Lopez (1996) and Teräsvirta (2008). There are also several comprehensive surveys that focus on the forecasting performance of GARCH models including Poon and Granger (2003), Poon (2005), and Andersen et al. (2006). However, there are relatively few surveys that focus on the practical econometric issues associated with estimating GARCH models and forecasting volatility. This chapter, which draws heavily from Zivot and Wang (2005), gives a tour through the empirical analysis of univariate GARCH models for financial time series with stops along the way to discuss various practical issues. Multivariate GARCH models are discussed in the chapter by Silvennoinen and Teräsvirta (2008). The plan of this chapter is as follows. Section 2 reviews some stylized facts of asset returns using example data on Microsoft and S&P 500 index returns. Section 3 reviews the basic univariate GARCH model. Testing for GARCH effects and estimation of GARCH models are covered in Sections 4 and 5. Asymmetric and non-Gaussian GARCH models are discussed in Section 6,

Eric Zivot
Department of Economics, University of Washington, Economics Box 353330, Seattle, WA 98195-3330, e-mail: ezivot@u.washington.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_5,

Asset	Mean	Med	Min	Max	Std. Dev	Skew	Kurt	JB
Daily Returns								
MSFT	0.0016	0.0000	-0.3012	0.1957	0.0253	-0.2457	11.66	13693
S&P 500	0.0004	0.0005	-0.2047	0.0909	0.0113	-1.486	32.59	160848
Monthly Returns								
MSFT	0.0336	0.0336	-0.3861	0.4384	0.1145	0.1845	4.004	9.922
S&P 500	0.0082	0.0122	-0.2066	0.1250	0.0459	-0.8377	5.186	65.75

Notes: Sample period is 03/14/86 - 06/30/03 giving 4365 daily observations.

Table 1 Summary Statistics for Daily and Monthly Stock Returns.

and long memory GARCH models are briefly discussed in Section 7. Section 8 discusses volatility forecasting, and final remarks are given Section 9 [1].

2 Some Stylized Facts of Asset Returns

Let P_t denote the price of an asset at the end of trading day t. The continuously compounded or log return is defined as $r_t = \ln(P_t/P_{t-1})$. Figure 1 plots the daily log returns, squared returns, and absolute value of returns of Microsoft stock and the S&P 500 index over the period March 14, 1986 through June 30, 2003. There is no clear discernible pattern of behavior in the log returns, but there is some persistence indicated in the plots of the squared and absolute returns which represent the volatility of returns. In particular, the plots show evidence of volatility clustering - low values of volatility followed by low values and high values of volatility followed by high values. This behavior is confirmed in Figure 2 which shows the sample autocorrelations of the six series. The log returns show no evidence of serial correlation, but the squared and absolute returns are positively autocorrelated. Also, the decay rates of the sample autocorrelations of r_t^2 and $|r_t|$ appear much slower, especially for the S&P 500 index, than the exponential rate of a covariance stationary autoregressive-moving average (ARMA) process suggesting possible long memory behavior. Monthly returns, defined as the sum of daily returns over the month, are illustrated in Figure 3. The monthly returns display much less volatility clustering than the daily returns.

Table 1 gives some standard summary statistics along with the Jarque-Bera test for normality. The latter is computed as

$$\text{JB} = \frac{T}{6}\left(\widehat{\text{skew}}^2 + \frac{(\widehat{\text{kurt}} - 3)^2}{4}\right), \tag{1}$$

[1] All of the examples in the paper were constructed using S-PLUS 8.0 and S+FinMetrics 2.0. Script files for replicating the examples may be downloaded from `http://faculty.washington.edu/ezivot`

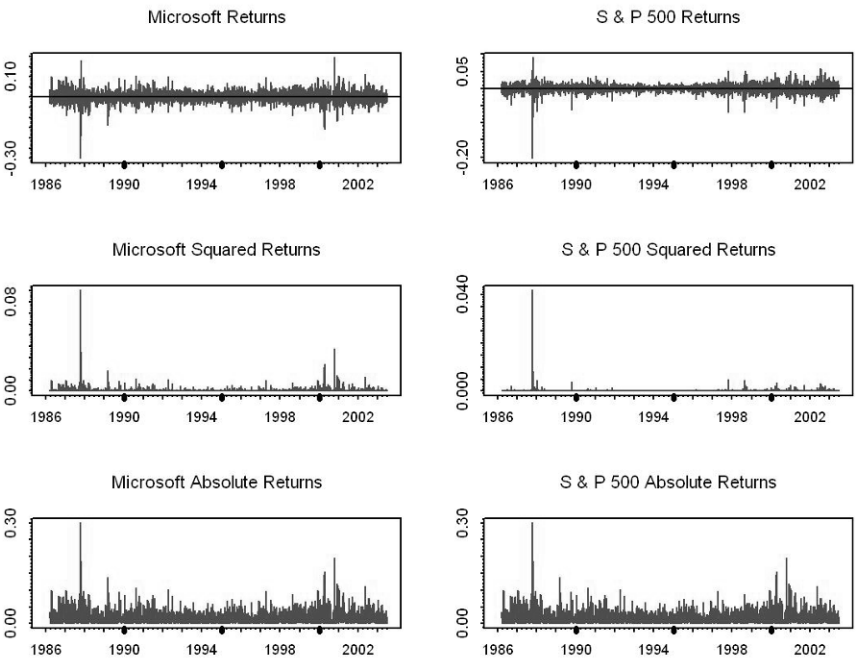

Fig. 1 Daily returns, squared returns and absolute returns for Microsoft and the S&P 500 index.

where $\widehat{\text{skew}}$ denotes the sample skewness and $\widehat{\text{kurt}}$ denotes the sample kurtosis. Under the null that the data are iid normal, JB is asymptotically distributed as chi-square with 2 degrees of freedom. The distribution of daily returns is clearly non-normal with negative skewness and pronounced excess kurtosis. Part of this non-normality is caused by some large outliers around the October 1987 stock market crash and during the bursting of the 2000 tech bubble. However, the distribution of the data still appears highly non-normal even after the removal of these outliers. Monthly returns have a distribution that is much closer to the normal than daily returns.

3 The ARCH and GARCH Model

Engle (1982) showed that the serial correlation in squared returns, or conditional heteroskedasticity, can be modeled using an autoregressive conditional heteroskedasticity (ARCH) model of the form

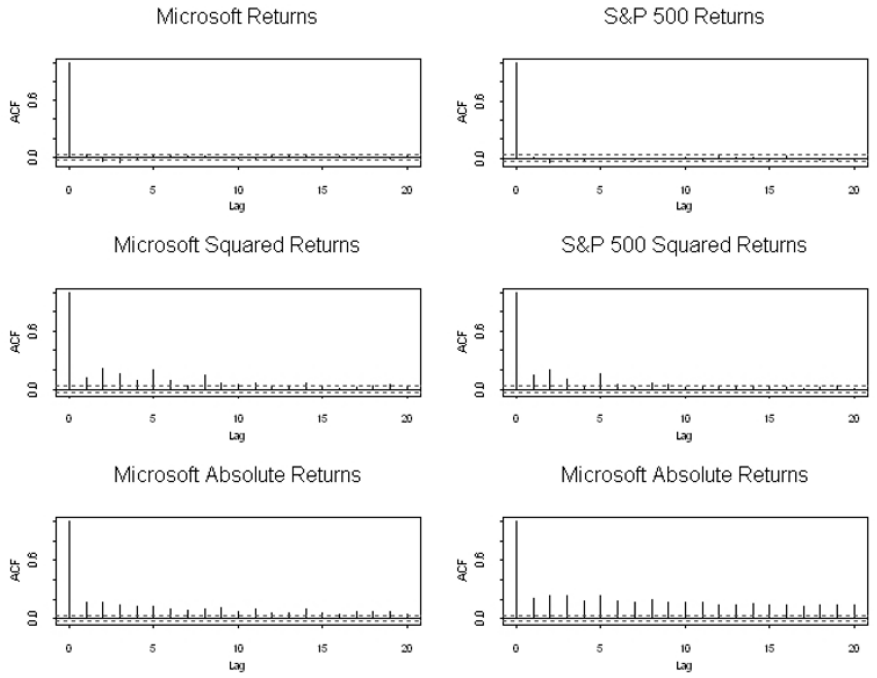

Fig. 2 Sample autocorrelations of r_t, r_t^2 and $|r_t|$ for Microsoft and S&P 500 index.

$$y_t = E_{t-1}[y_t] + \epsilon_t, \tag{2}$$

$$\epsilon_t = z_t \sigma_t, \tag{3}$$

$$\sigma_t^2 = a_0 + a_1 \epsilon_{t-1}^2 + \cdots + a_p \epsilon_{t-p}^2, \tag{4}$$

where $E_{t-1}[\cdot]$ represents expectation conditional on information available at time $t-1$, and z_t is a sequence of iid random variables with mean zero and unit variance. In the basic ARCH model z_t is assumed to be iid standard normal. The restrictions $a_0 > 0$ and $a_i \geq 0$ $(i = 1, \ldots, p)$ are required for $\sigma_t^2 > 0$. The representation (2) - (4) is convenient for deriving properties of the model as well as for specifying the likelihood function for estimation. The equation for σ_t^2 can be rewritten as an AR(p) process for ϵ_t^2

$$\epsilon_t^2 = a_0 + a_1 \epsilon_{t-1}^2 + \cdots + a_p \epsilon_{t-p}^2 + u_t, \tag{5}$$

where $u_t = \epsilon_t^2 - \sigma_t^2$ is a martingale difference sequence (MDS) since $E_{t-1}[u_t] = 0$ and it is assumed that $E(\epsilon_t^2) < \infty$. If $a_1 + \cdots + a_p < 1$ then ϵ_t is covariance stationary, the persistence of ϵ_t^2 and σ_t^2 is measured by $a_1 + \cdots + a_p$ and $\bar{\sigma}^2 = \text{var}(\epsilon_t) = E(\epsilon_t^2) = a_0/(1 - a_1 - \cdots - a_p)$.

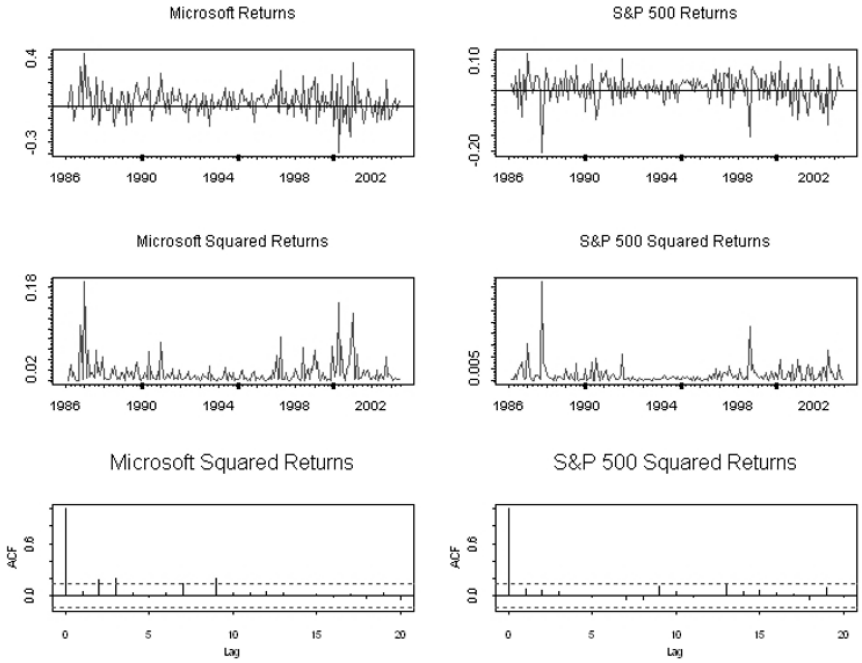

Fig. 3 Monthly Returns, Squared Returns and Sample Autocorrelations of Squared Returns for Microsoft and the S&P 500.

An important extension of the ARCH model proposed by Bollerslev (1986) replaces the AR(p) representation in (4) with an ARMA(p, q) formulation

$$\sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2, \tag{6}$$

where the coefficients a_i $(i = 0, \cdots, p)$ and b_j $(j = 1, \cdots, q)$ are all assumed to be positive to ensure that the conditional variance σ_t^2 is always positive.[2] The model in (6) together with (2)-(3) is known as the generalized ARCH or GARCH(p, q) model. The GARCH(p, q) model can be shown to be equivalent to a particular ARCH(∞) model. When $q = 0$, the GARCH model reduces to the ARCH model. In order for the GARCH parameters, b_j $(j = 1, \cdots, q)$, to be identified at least one of the ARCH coefficients a_i $(i > 0)$ must be nonzero. Usually a GARCH(1,1) model with only three parameters in the conditional variance equation is adequate to obtain a good model fit for financial time

[2] Positive coefficients are sufficient but not necessary conditions for the positivity of conditional variance. See Nelson (1992) and Conrad and Haag (2006) for more general conditions.

series. Indeed, Hansen and Lunde (2004) provided compelling evidence that is difficult to find a volatility model that outperforms the simple GARCH(1,1).

Just as an ARCH model can be expressed as an AR model of squared residuals, a GARCH model can be expressed as an ARMA model of squared residuals. Consider the GARCH(1,1) model

$$\sigma_t^2 = a_0 + a_1\epsilon_{t-1}^2 + b_1\sigma_{t-1}^2. \tag{7}$$

Since $E_{t-1}(\epsilon_t^2) = \sigma_t^2$, (7) can be rewritten as

$$\epsilon_t^2 = a_0 + (a_1 + b_1)\epsilon_{t-1}^2 + u_t - b_1 u_{t-1}, \tag{8}$$

which is an ARMA(1,1) model with $u_t = \epsilon_t^2 - E_{t-1}(\epsilon_t^2)$ being the MDS disturbance term.

Given the ARMA(1,1) representation of the GARCH(1,1) model, many of its properties follow easily from those of the corresponding ARMA(1,1) process for ϵ_t^2. For example, the persistence of σ_t^2 is captured by $a_1 + b_1$ and covariance stationarity requires that $a_1 + b_1 < 1$. The covariance stationary GARCH(1,1) model has an ARCH(∞) representation with $a_i = a_1 b_1^{i-1}$, and the unconditional variance of ϵ_t is $\bar{\sigma}^2 = a_0/(1 - a_1 - b_1)$.

For the general GARCH(p, q) model (6), the squared residuals ϵ_t behave like an ARMA($\max(p, q), q$) process. Covariance stationarity requires $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_i < 1$ and the unconditional variance of ϵ_t is

$$\bar{\sigma}^2 = \text{var}(\epsilon_t) = \frac{a_0}{1 - \left(\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_i\right)}. \tag{9}$$

3.1 Conditional mean specification

Depending on the frequency of the data and the type of asset, the conditional mean $E_{t-1}[y_t]$ is typically specified as a constant or possibly a low order autoregressive-moving average (ARMA) process to capture autocorrelation caused by market microstructure effects (e.g., bid-ask bounce) or non-trading effects. If extreme or unusual market events have happened during sample period, then dummy variables associated with these events are often added to the conditional mean specification to remove these effects. Therefore, the typical conditional mean specification is of the form

$$E_{t-1}[y_t] = c + \sum_{i=1}^{r} \phi_i y_{t-i} + \sum_{j=1}^{s} \theta_j \epsilon_{t-j} + \sum_{l=0}^{L} \beta_l' \mathbf{x}_{t-l} + \epsilon_t, \tag{10}$$

where $\mathbf{x}_t$ is a $k \times 1$ vector of exogenous explanatory variables.

In financial investment, high risk is often expected to lead to high returns. Although modern capital asset pricing theory does not imply such a simple relationship, it does suggest that there are some interactions between expected returns and risk as measured by volatility. Engle, Lilien and Robins (1987) proposed to extend the basic GARCH model so that the conditional volatility can generate a risk premium which is part of the expected returns. This extended GARCH model is often referred to as GARCH-in-the-mean or GARCH-M model . The GARCH-M model extends the conditional mean equation (10) to include the additional regressor $g(\sigma_t)$, which can be an arbitrary function of conditional volatility σ_t. The most common specifications are $g(\sigma_t) = \sigma_t^2$, σ_t, or $\ln(\sigma_t^2)$.

3.2 Explanatory variables in the conditional variance equation

Just as exogenous variables may be added to the conditional mean equation, exogenous explanatory variables may also be added to the conditional variance formula (6) in a straightforward way giving

$$\sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2 + \sum_{k=1}^{K} \delta_k' \mathbf{z}_{t-k},$$

where $\mathbf{z}_t$ is a $m \times 1$ vector of variables, and δ is a $m \times 1$ vector of positive coefficients. Variables that have been shown to help predict volatility are trading volume, macroeconomic news announcements (Lamoureux and Lastrapes (1990a), Flannery and Protopapadakis (2002), Bomfin (2003)), implied volatility from option prices and realized volatility (Taylor and Xu (1997), Blair et al. (2001)), overnight returns (Gallo and Pacini (1997), Martens (2002)), and after hours realized volatility (Chen et al. (2008))

3.3 The GARCH model and stylized facts of asset returns

Previously it was shown that the daily returns on Microsoft and the S&P 500 exhibited the "stylized facts" of volatility clustering as well as a non-normal empirical distribution. Researchers have documented these and many other stylized facts about the volatility of economic and financial time series. Bollerslev et al. (1994) gave a complete account of these facts. Using the ARMA representation of GARCH models shows that the GARCH model is capable

of explaining many of those stylized facts. The four most important ones are: volatility clustering, fat tails, volatility mean reversion, and asymmetry.

To understand volatility clustering, consider the GARCH$(1,1)$ model in (7). Usually the GARCH coefficient b_1 is found to be around 0.9 for many daily or weekly financial time series. Given this value of b_1, it is obvious that large values of σ^2_{t-1} will be followed by large values of σ^2_t, and small values of σ^2_{t-1} will be followed by small values of σ^2_t. The same reasoning can be obtained from the ARMA representation in (8), where large/small changes in ϵ^2_{t-1} will be followed by large/small changes in ϵ^2_t.

It is well known that the distribution of many high frequency financial time series usually have fatter tails than a normal distribution. That is, extreme values occur more often than implied by a normal distribution. Bollerslev (1986) gave the condition for the existence of the fourth order moment of a GARCH$(1,1)$ process. Assuming the fourth order moment exists, Bollerslev (1986) showed that the kurtosis implied by a GARCH$(1,1)$ process with normal errors is greater than 3, the kurtosis of a normal distribution. He and Teräsvirta (1999a) and He and Teräsvirta (1999b) extended these results to general GARCH(p,q) models. Thus a GARCH model with normal errors can replicate some of the fat-tailed behavior observed in financial time series. A more thorough discussion of extreme value theory for GARCH is given by Davis and Mikosch (2008). Most often a GARCH model with a non-normal error distribution is required to fully capture the observed fat-tailed behavior in returns. These models are reviewed in sub-Section 6.2.

Although financial markets may experience excessive volatility from time to time, it appears that volatility will eventually settle down to a long run level. Recall, the unconditional variance of ϵ_t for the stationary GARCH$(1,1)$ model is $\bar{\sigma}^2 = a_0/(1 - a_1 - b_1)$. To see that the volatility is always pulled toward this long run, the ARMA representation in (8) may be rewritten in mean-adjusted form as:

$$(\epsilon^2_t - \bar{\sigma}^2) = (a_1 + b_1)(\epsilon^2_{t-1} - \bar{\sigma}^2) + u_t - b_1 u_{t-1}. \tag{11}$$

If the above equation is iterated k times, it follows that

$$(\epsilon^2_{t+k} - \bar{\sigma}^2) = (a_1 + b_1)^k(\epsilon^2_t - \bar{\sigma}^2) + \eta_{t+k},$$

where η_t is a moving average process. Since $a_1 + b_1 < 1$ for a covariance stationary GARCH$(1,1)$ model, $(a_1 + b_1)^k \to 0$ as $k \to \infty$. Although at time t there may be a large deviation between ϵ^2_t and the long run variance, $\epsilon^2_{t+k} - \bar{\sigma}^2$ will approach zero "on average" as k gets large; i.e., the volatility "mean reverts" to its long run level $\bar{\sigma}^2$. The magnitude of $a_1 + b_1$ controls the speed of mean reversion. The so-called half-life of a volatility shock, defined as $\ln(0.5)/\ln(a_1 + b_1)$, measures the average time it takes for $|\epsilon^2_t - \bar{\sigma}^2|$ to decrease by one half. Obviously, the closer $a_1 + b_1$ is to one the longer is the half-life of a volatility shock. If $a_1 + b_1 > 1$, the GARCH model is non-

stationary and the volatility will eventually explode to infinity as $k \to \infty$. Similar arguments can be easily constructed for a GARCH(p, q) model.

The standard GARCH(p, q) model with Gaussian errors implies a symmetric distribution for y_t and so cannot account for the observed asymmetry in the distribution of returns. However, as shown in Section 6, asymmetry can easily be built into the GARCH model by allowing ϵ_t to have an asymmetric distribution or by explicitly modeling asymmetric behavior in the conditional variance equation (6).

3.4 Temporal aggregation

Volatility clustering and non-Gaussian behavior in financial returns is typically seen in weekly, daily or intraday data. The persistence of conditional volatility tends to increase with the sampling frequency[3]. However, as shown in Drost and Nijman (1993), for GARCH models there is no simple aggregation principle that links the parameters of the model at one sampling frequency to the parameters at another frequency. This occurs because GARCH models imply that the squared residual process follows an ARMA type process with MDS innovations which is not closed under temporal aggregation. The practical result is that GARCH models tend to be fit to the frequency at hand. This strategy, however, may not provide the best out-of-sample volatility forecasts. For example, Martens (2002) showed that a GARCH model fit to S&P 500 daily returns produces better forecasts of weekly and monthly volatility than GARCH models fit to weekly or monthly returns, respectively.

4 Testing for ARCH/GARCH Effects

The stylized fact of volatility clustering in returns manifests itself as autocorrelation in squared and absolute returns or in the residuals from the estimated conditional mean equation (10). The significance of these autocorrelations may be tested using the Ljung-Box or modified Q-statistic

$$\mathrm{MQ}(p) = T(T+2)\sum_{j=1}^{p}\frac{\hat{\rho}_j^2}{T-j}, \tag{12}$$

[3] The empirical result that aggregated returns exhibit smaller GARCH effects and approach Gaussian behavior can be explained by the results of Diebold (1988) who showed that a central limit theorem holds for standardized sums of random variables that follow covariance stationary GARCH processes.

where $\hat{\rho}_j$ denotes the j-lag sample autocorrelation of the squared or absolute returns. If the data are white noise then the MQ(p) statistic has an asymptotic chi-square distribution with p degrees of freedom. A significant value for MQ(p) provides evidence for time varying conditional volatility.

To test for autocorrelation in the raw returns when it is suspected that there are GARCH effects present, Diebold and Lopez (1996) suggested using the following heteroskedasticity robust version of (12)

$$\text{MQ}^{HC}(p) = T(T+2)\sum_{j=1}^{p}\frac{1}{T-j}\left(\frac{\hat{\sigma}^4}{\hat{\sigma}^4+\hat{\gamma}_j}\right)\hat{\rho}_j^2,$$

where $\hat{\sigma}^4$ is a consistent estimate of the squared unconditional variance of returns, and $\hat{\gamma}_j$ is the sample autocovariance of squared returns.

Since an ARCH model implies an AR model for the squared residuals ϵ_t^2, Engle (1982) showed that a simple Lagrange multiplier (LM) test for ARCH effects can be constructed based on the auxiliary regression (5). Under the null hypothesis that there are no ARCH effects, $a_1 = a_2 = \cdots = a_p = 0$, the test statistic

$$\text{LM} = T \cdot R^2 \tag{13}$$

has an asymptotic chi-square distribution with p degrees of freedom, where T is the sample size and R^2 is computed from the regression (5) using estimated residuals. Even though the LM test is constructed from an ARCH model, Lee and King (1993) show that it also has power against more general GARCH alternatives and so it can be used as a general specification test for GARCH effects.

Lumsdaine and Ng (1999), however, argued that the LM test (13) may reject if there is general misspecification in the conditional mean equation (10). They showed that such misspecification causes the estimated residuals $\hat{\epsilon}_t$ to be serially correlated which, in turn, causes $\hat{\epsilon}_t^2$ to be serially correlated. Therefore, care should be exercised in specifying the conditional mean equation (10) prior to testing for ARCH effects.

4.1 Testing for ARCH effects in daily and monthly returns

Table 2 shows values of MQ(p) computed from daily and monthly squared returns and the LM test for ARCH, for various values of p, for Microsoft and the S&P 500. There is clear evidence of volatility clustering in the daily returns, but less evidence for monthly returns especially for the S&P 500.

		MQ(p) r_t^2			LM		
Asset	p	1	5	10	1	5	10
				Daily Returns			
MSFT		56.81	562.1	206.8	56.76	377.9	416.6
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
S&P 500		87.59	415.5	456.1	87.52	311.4	329.8
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
				Monthly Returns			
MSFT		0.463	17.48	31.59	0.455	16.74	33.34
		(0.496)	(0.003)	(0.000)	(0.496)	(0.005)	(0.000)
S&P 500		1.296	2.590	6.344	1.273	2.229	5.931
		(0.255)	(0.763)	(0.786)	(0.259)	(0.817)	(0.821)

Notes: p-values are in parentheses.

Table 2 Tests for ARCH Effects in Daily Stock Returns

5 Estimation of GARCH Models

The general GARCH(p, q) model with normal errors is (2), (3) and (6) with $z_t \sim$ iid $N(0, 1)$. For simplicity, assume that $E_{t-1}[y_t] = c$. Given that ϵ_t follows Gaussian distribution conditional on past history, the prediction error decomposition of the log-likelihood function of the GARCH model conditional on initial values is

$$\log L = \sum_{t=1}^{T} l_t = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log\sigma_t^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{\epsilon_t^2}{\sigma_t^2}, \tag{14}$$

where $l_t = -\frac{1}{2}(\log(2\pi) + \log\sigma_t^2) - \frac{1}{2}\frac{\epsilon_t^2}{\sigma_t^2}$. The conditional loglikelihood (14) is used in practice since the unconditional distribution of the initial values is not known in closed form[4]. As discussed in McCullough and Renfro (1999) and Brooks et al. (2001), there are several practical issues to consider in the maximization of (14). Starting values for the model parameters c, a_i $(i = 0, \cdots, p)$ and b_j $(j = 1, \cdots, q)$ need to be chosen and an initialization of ϵ_t^2 and σ_t^2 must be supplied. The sample mean of y_t is usually used as the starting value for c, zero values are often given for the conditional variance parameters other than a_0 and a_1, and a_0 is set equal to the unconditional variance of y_t[5]. For the initial values of σ_t^2, a popular choice is

$$\sigma_t^2 = \epsilon_t^2 = \frac{1}{T}\sum_{s=1}^{T}\epsilon_s^2, \ t \leq 0,$$

[4] Diebold and Schuermann (1993) gave a computationally intensive numerical procedure for approximating the exact log-likelihood.

[5] Setting the starting values for all of the ARCH coefficients a_i $(i = 1, \ldots, p)$ to zero may create an ill-behaved likelihood and lead to a local minimum since the remaining GARCH parameters are not identified.

where the initial values for ϵ_s are computed as the residuals from a regression of y_t on a constant.

Once the log-likelihood is initialized, it can be maximized using numerical optimization techniques. The most common method is based on a Newton-Raphson iteration of the form

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \lambda_n \mathbf{H}(\hat{\theta}_n)^{-1}\mathbf{s}(\hat{\theta}_n),$$

where θ_n denotes the vector of estimated model parameters at iteration n, λ_n is a scalar step-length parameter, and $\mathbf{s}(\theta_n)$ and $\mathbf{H}(\theta_n)$ denote the gradient (or score) vector and Hessian matrix of the log-likelihood at iteration n, respectively. The step length parameter λ_n is chosen such that $\ln L(\theta_{n+1}) \geq \ln L(\theta_n)$. For GARCH models, the BHHH algorithm is often used. This algorithm approximates the Hessian matrix using only first derivative information

$$-\mathbf{H}(\theta) \approx \mathbf{B}(\theta) = \sum_{t=1}^{T} \frac{\partial l_t}{\partial \theta} \frac{\partial l_t}{\partial \theta'}.$$

In the application of the Newton-Raphson algorithm, analytic or numerical derivatives may be used. Fiorentini et al. (1996) provided algorithms for computing analytic derivatives for GARCH models.

The estimates that maximize the conditional log-likelihood (14) are called the maximum likelihood (ML) estimates. Under suitable regularity conditions, the ML estimates are consistent and asymptotically normally distributed and an estimate of the asymptotic covariance matrix of the ML estimates is constructed from an estimate of the final Hessian matrix from the optimization algorithm used. Unfortunately, verification of the appropriate regularity conditions has only been done for a limited number of simple GARCH models, see Lumsdaine (1992), Lee and Hansen (1993), Jensen and Rahbek (2004), Kristensen and Rahbek (2005) and Straumann (2005). In practice, it is generally assumed that the necessary regularity conditions are satisfied.

In GARCH models for which the distribution of z_t is symmetric and the parameters of the conditional mean and variance equations are variation free, the information matrix of the log-likelihood is block diagonal. The implication of this is that the parameters of the conditional mean equation can be estimated separately from those of the conditional variance equation without loss of asymptotic efficiency. This can greatly simplify estimation. An common model for which block diagonality of the information matrix fails is the GARCH-M model.

5.1 Numerical accuracy of GARCH estimates

GARCH estimation is widely available in a number of commercial software packages (e.g. EVIEWS, GAUSS, MATLAB, Ox, RATS, S-PLUS, TSP) and there are also a few free open source implementations. Fiorentini et al. (1996), McCullough and Renfro (1999), and Brooks et al. (2001) discussed numerical accuracy issues associated with maximizing the GARCH log-likelihood. They found that starting values, optimization algorithm choice, and use of analytic or numerical derivatives, and convergence criteria all influence the resulting numerical estimates of the GARCH parameters. McCullough and Renfro (1999) and Brooks et al. (2001) studied estimation of a GARCH(1,1) model from a variety of commercial statistical packages using the exchange rate data of Bollerslev and Ghysels (1996) as a benchmark. They found that it is often difficult to compare competing software since the exact construction of the GARCH likelihood is not always adequately described. In general, they found that use of analytic derivatives leads to more accurate estimation than procedures based on purely numerical evaluations.

In practice, the GARCH log-likelihood function is not always well behaved, especially in complicated models with many parameters, and reaching a global maximum of the log-likelihood function is not guaranteed using standard optimization techniques. Also, the positive variance and stationarity constraints are not straightforward to implement with common optimization software and are often ignored in practice. Poor choice of starting values can lead to an ill-behaved log-likelihood and cause convergence problems. Therefore, it is always a good idea to explore the surface of the log-likelihood by perturbing the starting values and re-estimating the GARCH parameters.

In many empirical applications of the GARCH(1,1) model, the estimate of a_1 is close to zero and the estimate of b_1 is close to unity. This situation is of some concern since the GARCH parameter b_1 becomes unidentified if $a_1 = 0$, and it is well known that the distribution of ML estimates can become ill-behaved in models with nearly unidentified parameters. Ma et al. (2007) studied the accuracy of ML estimates of the GARCH parameters a_0, a_1 and b_1 when a_1 is close to zero. They found that the estimated standard error for b_1 is spuriously small and that the t-statistics for testing hypotheses about the true value of b_1 are severely size distorted. They also showed that the concentrated loglikelihood as a function of b_1 exhibits multiple maxima. To guard against spurious inference they recommended comparing estimates from pure ARCH(p) models, which do not suffer from the identification problem, with estimates from the GARCH(1,1). If the volatility dynamics from these models are similar then the spurious inference problem is not likely to be present.

5.2 Quasi-maximum likelihood estimation

Another practical issue associated with GARCH estimation concerns the correct choice of the error distribution. In particular, the assumption of conditional normality is not always appropriate. However, as shown by Weiss (1986) and Bollerslev and Woolridge (1992), even when normality is inappropriately assumed, maximizing the Gaussian log-likelihood (14) results in quasi-maximum likelihood estimates (QMLEs) that are consistent and asymptotically normally distributed provided the conditional mean and variance functions of the GARCH model are correctly specified. In addition, Bollerslev and Woolridge (1992) derived an asymptotic covariance matrix for the QMLEs that is robust to conditional non-normality. This matrix is estimated using

$$\mathbf{H}(\hat{\theta}_{QML})^{-1}\mathbf{B}(\hat{\theta}_{QML})\mathbf{H}(\hat{\theta}_{QML})^{-1}, \tag{15}$$

where $\hat{\theta}_{QML}$ denotes the QMLE of θ, and is often called the "sandwich" estimator. The coefficient standard errors computed from the square roots of the diagonal elements of (15) are sometimes called "Bollerslev-Wooldridge" standard errors. Of course, the QMLEs will be less efficient than the true MLEs based on the correct error distribution. However, if the normality assumption is correct then the sandwich covariance is asymptotically equivalent to the inverse of the Hessian. As a result, it is good practice to routinely use the sandwich covariance for inference purposes.

Engle and González-Rivera (1991) and Bollerslev and Woolridge (1992) evaluated the accuracy of the quasi-maximum likelihood estimation (QMLE) of GARCH(1,1) models. They found that if the distribution of z_t in (3) is symmetric, then QMLE is often close to the MLE. However, if z_t has a skewed distribution then the QMLE can be quite different from the MLE.

A detailed description of the asymptotic theory of GARCH estimation can be found in Francq and Zakoïan (2008).

5.3 Model selection

An important practical problem is the determination of the ARCH order p and the GARCH order q for a particular series. Since GARCH models can be treated as ARMA models for squared residuals, traditional model selection criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) can be used for selecting models. For daily returns, if attention is restricted to pure ARCH(p) models it is typically found that large values of p are selected by AIC and BIC. For GARCH(p, q) models, those with $p, q \leq 2$ are typically selected by AIC and BIC. Low order GARCH(p,q) models are generally preferred to a high order ARCH(p) for reasons of parsimony and better numerical stability of estimation (high

order GARCH(p, q) processes often have many local maxima and minima). For many applications, it is hard to beat the simple GARCH(1,1) model.

For more details on general model selection procedures and its properties it is referred to Leeb and Pötscher (2008).

5.4 Evaluation of estimated GARCH models

After a GARCH model has been fit to the data, the adequacy of the fit can be evaluated using a number of graphical and statistical diagnostics. If the GARCH model is correctly specified, then the estimated standardized residuals $\hat{\epsilon}_t/\hat{\sigma}_t$ should behave like classical regression residuals; i.e., they should not display serial correlation, conditional heteroskedasticity or any type of nonlinear dependence. In addition, the distribution of the standardized residuals $\hat{\epsilon}_t/\hat{\sigma}_t$ should match the specified error distribution used in the estimation.

Graphically, ARCH effects reflected by serial correlation in $\hat{\epsilon}_t^2/\hat{\sigma}_t^2$ can be uncovered by plotting its SACF. The modified Ljung-Box statistic (12) can be used to test the null of no autocorrelation up to a specific lag, and Engle's LM statistic (13) can be used to test the null of no remaining ARCH effects[6]. If it is assumed that the errors are Gaussian, then a plot of $\hat{\epsilon}_t/\hat{\sigma}_t$ against time should have roughly ninety five percent of its values between ± 2; a normal qq-plot of $\hat{\epsilon}_t/\hat{\sigma}_t$ should look roughly linear[7]; and the JB statistic should not be too much larger than six.

5.5 Estimation of GARCH models for daily and monthly returns

Table 3 gives model selection criteria for a variety of GARCH(p, q) fitted to the daily returns on Microsoft and the S&P 500. For pure ARCH(p) models, an ARCH(5) is chosen by all criteria for both series. For GARCH(p, q) models, AIC picks a GARCH(2,1) for both series and BIC picks a GARCH(1,1) for both series[8].

Table 4 gives QMLEs of the GARCH(1,1) model assuming normal errors for the Microsoft and S&P 500 daily returns. For both series, the estimates

[6] These tests should be viewed as indicative, since the distribution of the tests are influenced by the estimation of the GARCH model. For valid LM tests, the partial derivatives of σ_t^2 with respect to the conditional volatility parameters should be added as additional regressors in the auxiliary regression (5) based on estimated residuals.

[7] If an error distribution other than the Gaussian is assumed, then the qq-plot should be constructed using the quantiles of the assumed distribution.

[8] The low log-likelihood values for the GARCH(2,2) models indicate that a local maximum was reached.

(p,q)	Asset	AIC	BIC	Likelihood
(1,0)	MSFT	-19977	-19958	9992
	S&P 500	-27337	-27318	13671
(2,0)	MSFT	-20086	-20060	10047
	S&P 500	-27584	-27558	13796
(3,0)	MSFT	-20175	-20143	10092
	S&P 500	-27713	-27681	13861
(4,0)	MSFT	-20196	-20158	10104
	S&P 500	-27883	-27845	13947
(5,0)	MSFT	-20211	-20166	10113
	S&P 500	-27932	-27887	13973
(1,1)	MSFT	-20290	-20264	10149
	S&P 500	-28134	-28109	14071
(1,2)	MSFT	-20290	-20258	10150
	S&P 500	-28135	-28103	14072
(2,1)	MSFT	-20292	-20260	10151
	S&P 500	-28140	-28108	14075
(2,2)	MSFT	-20288	-20249	10150
	S&P 500	-27858	-27820	13935

Table 3 Model Selection Criteria for Estimated GARCH(p,q) Models.

of a_1 are around 0.09 and the estimates of b_1 are around 0.9. Using both ML and QML standard errors, these estimates are statistically different from zero. However, the QML standard errors are considerably larger than the ML standard errors. The estimated volatility persistence, $a_1 + b_1$, is very high for both series and implies half-lives of shocks to volatility to Microsoft and the S&P 500 of 15.5 days and 76 days, respectively. The unconditional standard deviation of returns, $\bar{\sigma} = \sqrt{a_0/(1 - a_1 - b_1)}$, for Microsoft and the S&P 500 implied by the GARCH(1,1) models are 0.0253 and 0.0138, respectively, and are very close to the sample standard deviations of returns reported in Table 1.

Estimates of GARCH-M(1,1) models for Microsoft and the S&P 500, where σ_t is added as a regressor to the mean equation, show small positive coefficients on σ_t and essentially the same estimates for the remaining parameters as the GARCH(1,1) models.

Figure 4 shows the first differences of returns along with the fitted one-step-ahead volatilities, $\hat{\sigma}_t$, computed from the GARCH(1,1) and ARCH(5) models. The ARCH(5) and GARCH(1,1) models do a good job of capturing the observed volatility clustering in returns. The GARCH(1,1) volatilities, however, are smoother and display more persistence than the ARCH(5) volatilities.

Graphical diagnostics from the fitted GARCH(1,1) models are illustrated in Figure 5. The SACF of $\hat{\epsilon}_t^2/\hat{\sigma}_t^2$ does not indicate any significant autocorrelation, but the normal qq-plot of $\hat{\epsilon}_t/\hat{\sigma}_t$ shows strong departures from normality. The last three columns of Table 4 give the standard statistical diagnostics of the fitted GARCH models. Consistent with the SACF, the MQ statistic and

	GARCH Parameters			Residual Diagnostics		
Asset	a_0	a_1	b_1	MQ(12)	LM(12)	JB
			Daily Returns			
MSFT	$2.80e^{-5}$ $(3.42e^{-6})$ $[1.10e^{-5}]$	0.0904 (0.0059) [0.0245]	0.8658 (0.0102) [0.0371]	4.787 (0.965)	4.764 (0.965)	1751 (0.000)
S&P 500	$1.72e^{-6}$ $(2.00e^{-7})$ $[1.25e^{-6}]$	0.0919 (0.0029) [0.0041]	0.8990 (0.0046) [0.0436]	5.154 (0.953)	5.082 (0.955)	5067 (0.000)
			Monthly Returns			
MSFT	0.0006 [0.0006]	0.1004 [0.0614]	0.8525 [0.0869]	8.649 (0.733)	6.643 (0.880)	3.587 (0.167)
S&P 500	$3.7e^{-5}$ $[9.6e^{-5}]$	0.0675 [0.0248]	0.9179 [0.0490]	3.594 (0.000)	3.660 (0.988)	72.05 (0.000)

Notes: QML standard errors are in brackets.

Table 4 Estimates of GARCH(1,1) Model with Diagnostics.

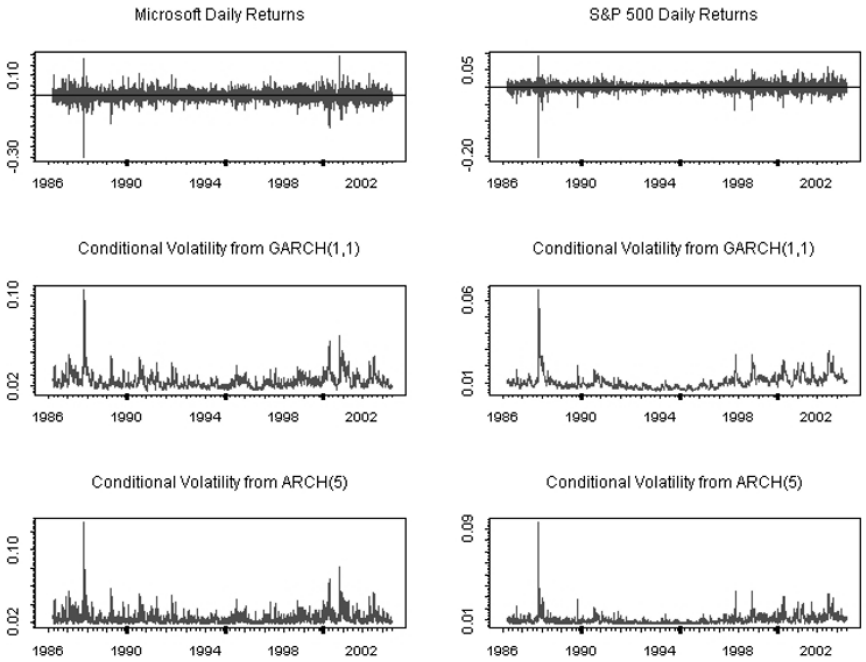

Fig. 4 One-step ahead volatilities from fitted ARCH(5) and GARCH(1,1) models for Microsoft and S&P 500 index.

Engle's LM statistic do not indicate remaining ARCH effects. Furthermore, the extremely large JB statistic confirms nonnormality.

Table 4 also shows estimates of GARCH(1,1) models fit to the monthly returns. The GARCH(1,1) models fit to the monthly returns are remarkable similar to those fit to the daily returns. There are, however, some important differences. The monthly standardized residuals are much closer to the normal distribution, especially for Microsoft. Also, the GARCH estimates for the S&P 500 reflect some of the characteristics of spurious GARCH effects as discussed in Ma et al. (2007). In particular, the estimate of a_1 is close to zero, and has a relatively large QML standard error, and the estimate of b_1 is close to one and has a very small standard error.

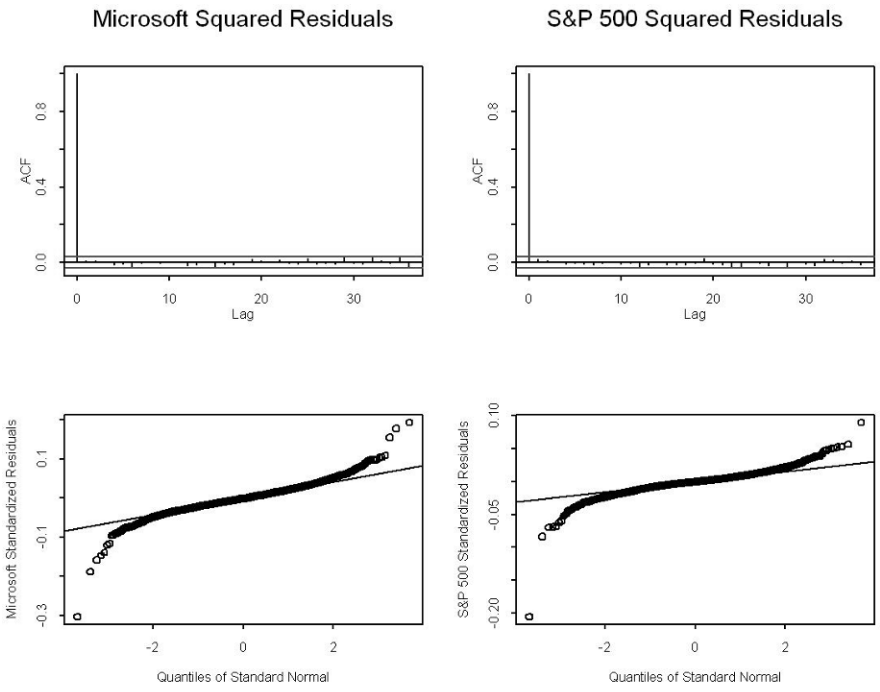

Fig. 5 Graphical residual diagnostics from fitted GARCH(1,1) models to Microsoft and S&P 500 returns.

6 GARCH Model Extensions

In many cases, the basic GARCH conditional variance equation (6) under normality provides a reasonably good model for analyzing financial time series and estimating conditional volatility. However, in some cases there are aspects of the model which can be improved so that it can better capture the characteristics and dynamics of a particular time series. For example, the empirical analysis in the previous Section showed that for the daily returns on Microsoft and the S&P 500, the normality assumption may not be appropriate and there is evidence of nonlinear behavior in the standardized residuals from the fitted GARCH(1,1) model. This Section discusses several extensions to the basic GARCH model that make GARCH modeling more flexible.

6.1 Asymmetric leverage effects and news impact

In the basic GARCH model (6), since only squared residuals ϵ_{t-i}^2 enter the conditional variance equation, the signs of the residuals or shocks have no effect on conditional volatility. However, a stylized fact of financial volatility is that bad news (negative shocks) tends to have a larger impact on volatility than good news (positive shocks). That is, volatility tends to be higher in a falling market than in a rising market. Black (1976) attributed this effect to the fact that bad news tends to drive down the stock price, thus increasing the leverage (i.e., the debt-equity ratio) of the stock and causing the stock to be more volatile. Based on this conjecture, the asymmetric news impact on volatility is commonly referred to as the leverage effect.

6.1.1 Testing for asymmetric effects on conditional volatility

A simple diagnostic for uncovering possible asymmetric leverage effects is the sample correlation between r_t^2 and r_{t-1}. A negative value of this correlation provides some evidence for potential leverage effects. Other simple diagnostics, suggested by Engle and Ng (1993), result from estimating the following test regression

$$\hat{\varepsilon}_t^2 = \beta_0 + \beta_1 \hat{w}_{t-1} + \xi_t,$$

where $\hat{\varepsilon}_t$ is the estimated residual from the conditional mean equation (10), and $\hat{w}_{t-1}$ is a variable constructed from $\hat{\varepsilon}_{t-1}$ and the sign of $\hat{\varepsilon}_{t-1}$. A significant value of β_1 indicates evidence for asymmetric effects on conditional volatility. Let S_{t-1}^- denote a dummy variable equal to unity when $\hat{\varepsilon}_{t-1}$ is negative, and zero otherwise. Engle and Ng consider three tests for asymmetry. Setting $\hat{w}_{t-1} = S_{t-1}^-$ gives the Sign Bias test; setting $\hat{w}_{t-1} = S_{t-1}^- \hat{\varepsilon}_{t-1}$ gives the

Negative Size Bias test; and setting $\hat{w}_{t-1} = S^{+}_{t-1}\hat{\varepsilon}_{t-1}$ gives the Positive Size Bias test.

6.1.2 Asymmetric GARCH models

The leverage effect can be incorporated into a GARCH model in several ways. Nelson (1991) proposed the following exponential GARCH (EGARCH) model to allow for leverage effects

$$h_t = a_0 + \sum_{i=1}^{p} a_i \frac{|\epsilon_{t-i}| + \gamma_i \epsilon_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^{q} b_j h_{t-j}, \tag{16}$$

where $h_t = \log \sigma_t^2$. Note that when ϵ_{t-i} is positive or there is "good news", the total effect of ϵ_{t-i} is $(1+\gamma_i)|\epsilon_{t-i}|$; in contrast, when ϵ_{t-i} is negative or there is "bad news", the total effect of ϵ_{t-i} is $(1-\gamma_i)|\epsilon_{t-i}|$. Bad news can have a larger impact on volatility, and the value of γ_i would be expected to be negative. An advantage of the EGARCH model over the basic GARCH model is that the conditional variance σ_t^2 is guaranteed to be positive regardless of the values of the coefficients in (16), because the logarithm of σ_t^2 instead of σ_t^2 itself is modeled. Also, the EGARCH is covariance stationary provided $\sum_{j=1}^{q} b_j < 1$.

Another GARCH variant that is capable of modeling leverage effects is the threshold GARCH (TGARCH) model, [9] which has the following form

$$\sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \gamma_i S_{t-i} \epsilon_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2, \tag{17}$$

where

$$S_{t-i} = \begin{cases} 1 & \text{if } \epsilon_{t-i} < 0 \\ 0 & \text{if } \epsilon_{t-i} \geq 0 \end{cases}.$$

That is, depending on whether ϵ_{t-i} is above or below the threshold value of zero, ϵ_{t-i}^2 has different effects on the conditional variance σ_t^2: when ϵ_{t-i} is positive, the total effects are given by $a_i \epsilon_{t-i}^2$; when ϵ_{t-i} is negative, the total effects are given by $(a_i + \gamma_i)\epsilon_{t-i}^2$. So one would expect γ_i to be positive for bad news to have larger impacts.

Ding et al. (1993) extended the basic GARCH model to allow for leverage effects. Their *power* GARCH (PGARCH(p, d, q)) model has the form

$$\sigma_t^d = a_0 + \sum_{i=1}^{p} a_i (|\epsilon_{t-i}| + \gamma_i \epsilon_{t-i})^d + \sum_{j=1}^{q} b_j \sigma_{t-j}^d, \tag{18}$$

[9] The original TGARCH model proposed by Zakoïan (1994) models σ_t instead of σ_t^2. The TGARCH model is also known as the GJR model because Glosten et al. (1993) proposed essentially the same model.

GARCH(1, 1)	$\sigma_t^2 = A + a_1(\lvert\epsilon_{t-1}\rvert + \gamma_1\epsilon_{t-1})^2$ $A = a_0 + b_1\bar{\sigma}^2$ $\bar{\sigma}^2 = a_0/[1 - a_1(1 + \gamma_1^2) - b_1]$
TGARCH(1, 1)	$\sigma_t^2 = A + (a_1 + \gamma_1 S_{t-1})\epsilon_{t-1}^2$ $A = a_0 + b_1\bar{\sigma}^2$ $\bar{\sigma}^2 = a_0/[1 - (a_1 + \gamma_1/2) - b_1]$
PGARCH(1, 1, 1)	$\sigma_t^2 = A + 2\sqrt{A}a_1(\lvert\epsilon_{t-1}\rvert + \gamma_1\epsilon_{t-1})$ $+a_1^2(\lvert\epsilon_{t-1}\rvert + \gamma_1\epsilon_{t-1})^2$, $A = (a_0 + b_1\bar{\sigma})^2$ $\bar{\sigma}^2 = a_0^2/[1 - a_1/\sqrt{2/\pi} - b_1]^2$
EGARCH(1, 1)	$\sigma_t^2 = A\exp\{a_1(\lvert\epsilon_{t-1}\rvert + \gamma_1\epsilon_{t-1})/\bar{\sigma}\}$ $A = \bar{\sigma}^{2b_1}\exp\{a_0\}$ $\bar{\sigma}^2 = \exp\{(a_0 + a_1\sqrt{2/\pi})/(1 - b_1)\}$

Table 5 News impact curves for asymmetric GARCH processes. $\bar{\sigma}^2$ denotes the unconditional variance.

where d is a positive exponent, and γ_i denotes the coefficient of leverage effects. When $d = 2$, (18) reduces to the basic GARCH model with leverage effects. When $d = 1$, the PGARCH model is specified in terms of σ_t which tends to be less sensitive to outliers than when $d = 2$. The exponent d may also be estimated as an additional parameter which increases the flexibility of the model. Ding et al. (1993) showed that the PGARCH model also includes many other GARCH variants as special cases.

Many other asymmetric GARCH models have been proposed based on smooth transition and Markov switching models. See Franses and van Dijk (2000) and Teräsvirta (2008) for excellent surveys of these models.

6.1.3 News impact curve

The GARCH, EGARCH, TGARCH and PGARCH models are all capable of modeling leverage effects. To clearly see the impact of leverage effects in these models, Pagan and Schwert (1990), and Engle and Ng (1993) advocated the use of the so-called news impact curve. They defined the news impact curve as the functional relationship between conditional variance at time t and the shock term (error term) at time $t-1$, holding constant the information dated $t-2$ and earlier, and with all lagged conditional variance evaluated at the level of the unconditional variance. Table 5 summarizes the expressions defining the news impact curves, which include expressions for the unconditional variances, for the asymmetric GARCH(1,1) models.

Asset	$\text{corr}(r_t^2, r_{t-1})$	Sign Bias	Negative Size Bias	Positive Size Bias
Microsoft	−0.0315	−0.4417	−6.816	3.174
		(0.6587)	(0.000)	(0.001)
S&P 500	−0.098	2.457	−11.185	1.356
		(0.014)	(0.000)	(0.175)

Notes: p-values are in parentheses.

Table 6 Tests for Asymmetric GARCH Effects.

Model	a_0	a_1	b_1	γ_1	BIC
			Microsoft		
EGARCH	−0.7273	0.2144	0.9247	−0.2417	-20265
	[0.4064]	[0.0594]	[0.0489]	[0.0758]	
TGARCH	$3.01e^{-5}$	0.0564	0.8581	0.0771	-20291
	$[1.02e^{-5}]$	[0.0141]	[0.0342]	[0.0306]	
PGARCH 2	$2.87e^{-5}$	0.0853	0.8672	−0.2164	-20290
	$[9.27e^{-6}]$	[0.0206]	[0.0313]	[0.0579]	
PGARCH 1	0.0010	0.0921	0.8876	−0.2397	-20268
	[0.0006]	[0.0236]	[0.0401]	[0.0813]	
			S&P 500		
EGARCH	−0.2602	0.0720	0.9781	−0.3985	-28051
	[0.3699]	[0.0397]	[0.0389]	[0.4607]	
TGARCH	$1.7e^{-6}$	0.0157	0.9169	0.1056	-28200
	$[7.93e^{-7}]$	[0.0081]	[0.0239]	0.0357	
PGARCH 2	$1.78e^{-6}$	0.0578	0.9138	−0.4783	-28202
	$[8.74e^{-7}]$	[0.0165]	[0.0253]	[0.0910]	
PGARCH 1	0.0002	0.0723	0.9251	−0.7290	-28253
	$[2.56e^{-6}]$	[0.0003]	$[8.26e^{-6}]$	[0.0020]	

Notes: QML standard errors are in brackets.

Table 7 Estimates of Asymmetric GARCH(1,1) Models.

6.1.4 Asymmetric GARCH models for daily returns

Table 6 shows diagnostics and tests for asymmetric effects in the daily returns on Microsoft and the S&P 500. The correlation between r_t^2 and r_{t-1} is negative and fairly small for both series indicating weak evidence for asymmetry. However, the Size Bias tests clearly indicate asymmetric effects with the Negative Size Bias test giving the most significant results.

Table 7 gives the estimation results for EGARCH(1,1), TGARCH(1,1) and PGARCH(1,d,1) models for $d = 1, 2$. All of the asymmetric models show statistically significant leverage effects, and lower BIC values than the symmetric GARCH models. Model selection criteria indicate that the TGARCH(1,1) is the best fitting model for Microsoft, and the PGARCH(1,1,1) is the best fitting model for the S&P 500.

Figure 6 shows the estimated news impact curves based on these models. In this plot, the range of ϵ_t is determined by the residuals from the fitted models. The TGARCH and PGARCH(1,2,1) models have very similar NICs and show much larger responses to negative shocks than to positive shocks.

Since the EGARCH(1,1) and PGARCH(1,1,1) models are more robust to extreme shocks, impacts of small (large) shocks for these model are larger (smaller) compared to those from the other models and the leverage effect is less pronounced.

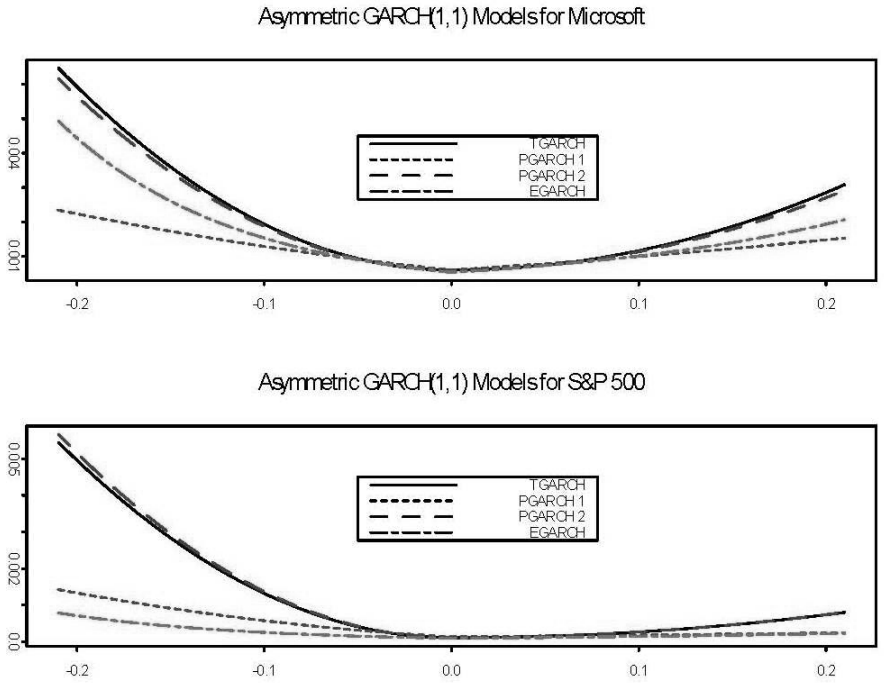

Fig. 6 News impact curves from fitted asymmetric GARCH(1,1) models for Microsoft and S&P 500 index.

6.2 Non-Gaussian error distributions

In all the examples illustrated so far, a normal error distribution has been exclusively used. However, given the well known fat tails in financial time series, it may be more appropriate to use a distribution which has fatter tails than the normal distribution. The most common fat-tailed error distributions for fitting GARCH models are: the Student's t distribution; the double exponential distribution; and the generalized error distribution.

Bollerslev (1987) proposed fitting a GARCH model with a Student's t distribution for the standardized residual. If a random variable u_t has a Stu-

dent's t distribution with ν degrees of freedom and a scale parameter s_t, the probability density function (pdf) of u_t is given by

$$f(u_t) = \frac{\Gamma[(\nu+1)/2]}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \frac{s_t^{-1/2}}{[1+u_t^2/(s_t\nu)]^{(\nu+1)/2}},$$

where $\Gamma(\cdot)$ is the gamma function. The variance of u_t is given by

$$\text{var}(u_t) = \frac{s_t\nu}{\nu-2},\ v > 2.$$

If the error term ϵ_t in a GARCH model follows a Student's t distribution with ν degrees of freedom and $\text{var}_{t-1}(\epsilon_t) = \sigma_t^2$, the scale parameter s_t should be chosen to be

$$s_t = \frac{\sigma_t^2(\nu-2)}{\nu}.$$

Thus the log-likelihood function of a GARCH model with Student's t distributed errors can be easily constructed based on the above pdf.

Nelson (1991) proposed to use the generalized error distribution (GED) to capture the fat tails usually observed in the distribution of financial time series. If a random variable u_t has a GED with mean zero and unit variance, the pdf of u_t is given by

$$f(u_t) = \frac{\nu\exp[-(1/2)|u_t/\lambda|^\nu]}{\lambda\cdot 2^{(\nu+1)/\nu}\Gamma(1/\nu)},$$

where

$$\lambda = \left[\frac{2^{-2/\nu}\Gamma(1/\nu)}{\Gamma(3/\nu)}\right]^{1/2},$$

and ν is a positive parameter governing the thickness of the tail behavior of the distribution. When $\nu = 2$ the above pdf reduces to the standard normal pdf; when $\nu < 2$, the density has thicker tails than the normal density; when $\nu > 2$, the density has thinner tails than the normal density.

When the tail thickness parameter $\nu = 1$, the pdf of GED reduces to the pdf of double exponential distribution:

$$f(u_t) = \frac{1}{\sqrt{2}}e^{-\sqrt{2}|u_t|}.$$

Based on the above pdf, the log-likelihood function of GARCH models with GED or double exponential distributed errors can be easily constructed. See to Hamilton (1994) for an example.

Several other non-Gaussian error distribution have been proposed. Fernandez and Steel (1998) introduced the asymmetric Student's t distribution to capture both skewness and excess kurtosis in the standardized residuals. Venter and de Jongh (2002) proposed the normal inverse Gaussian distribu-

Model	a_0	a_1	b_1	γ_1	v	BIC
			Microsoft			
GARCH	$3.39e^{-5}$	0.0939	0.8506		6.856	-20504
	$[1.52e^{-5}]$	[0.0241]	[0.0468]		[0.7121	
TGARCH	$3.44e^{-5}$	0.0613	0.8454	0.0769	7.070	-20511
	$[1.20e^{-5}]$	[0.0143]	[0.0380]	[0.0241]	[0.7023]	
			S&P 500			
GARCH	$5.41e^{-7}$	0.0540	0.0943		5.677	-28463
	$[2.15e^{-7}]$	[0.0095]	[0.0097]		[0.5571]	
PGARCH	0.0001	0.0624	0.9408	−0.7035	6.214	-28540
$d = 1$	[0.0002]	[0.0459]	[0.0564]	[0.0793]	[0.6369]	

Notes: QML standard errors are in brackets.

Table 8 Estimates of Non Gaussian GARCH(1,1) Models.

tion. Gallant and Tauchen (2001) provided a very flexible seminonparametric innovation distribution based on a Hermite expansion of a Gaussian density. Their expansion is capable of capturing general shape departures from Gaussian behavior in the standardized residuals of the GARCH model.

6.2.1 Non-Gaussian GARCH models for daily returns

Table 8 gives estimates of the GARCH(1,1) and best fitting asymmetric GARCH(1,1) models using Student's t innovations for the Microsoft and S&P 500 returns. Model selection criteria indicated that models using the Student's t distribution fit better than the models using the GED distribution. The estimated degrees of freedom for Microsoft is about 7, and for the S&P 500 about 6. The use of t-distributed errors clearly improves the fit of the GARCH(1,1) models. Indeed, the BIC values are even lower than the values for the asymmetric GARCH(1,1) models based on Gaussian errors (see Table 7). Overall, the asymmetric GARCH(1,1) models with t-distributed errors are the best fitting models. The qq-plots in Figure 7 shows that the Student's t distribution adequately captures the fat-tailed behavior in the standardized residuals for Microsoft but not for the S&P 500 index.

7 Long Memory GARCH Models

If returns follow a GARCH(p, q) model, then the autocorrelations of the squared and absolute returns should decay exponentially. However, the SACF of r_t^2 and $|r_t|$ for Microsoft and the S&P 500 in Figure 2 appear to decay much more slowly. This is evidence of so-called *long memory* behavior. Formally, a stationary process has long memory or long range dependence if its autocorrelation function behaves like

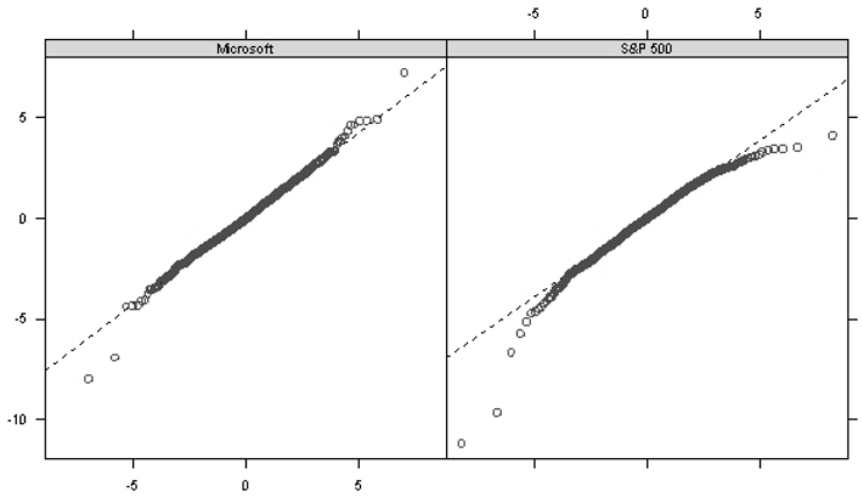

Fig. 7 QQ-plots of Standardized Residuals from Asymmetric GARCH(1,1) models with Student's t errors.

$$\rho(k) \rightarrow C_\rho k^{2d-1} \text{ as } k \rightarrow \infty,$$

where C_ρ is a positive constant, and d is a real number between 0 and $\frac{1}{2}$. Thus the autocorrelation function of a long memory process decays slowly at a hyperbolic rate. In fact, it decays so slowly that the autocorrelations are not summable:

$$\sum_{k=-\infty}^{\infty} \rho(k) = \infty.$$

It is important to note that the scaling property of the autocorrelation function does not dictate the general behavior of the autocorrelation function. Instead, it only specifies the asymptotic behavior when $k \rightarrow \infty$. What this means is that for a long memory process, it is not necessary for the autocorrelation to remain significant at large lags as long as the autocorrelation function decays slowly. Beran (1994) gives an example to illustrate this property.

The following subsections describe testing for long memory and GARCH models that can capture long memory behavior in volatility. Explicit long memory GARCH models are discussed in Teräsvirta (2008).

7.1 Testing for long memory

One of the best-known and easiest to use tests for long memory or long range dependence is the rescaled range (R/S) statistic, which was originally proposed by Hurst (1951), and later refined by Mandelbrot (1975) and his coauthors. The R/S statistic is the range of partial sums of deviations of a time series from its mean, rescaled by its standard deviation. Specifically, consider a time series y_t, for $t = 1, \cdots, T$. The R/S statistic is defined as

$$Q_T = \frac{1}{s_T}\left[\max_{1\le k\le T}\sum_{j=1}^{k}(y_j - \bar{y}) - \min_{1\le k\le T}\sum_{j=1}^{k}(y_j - \bar{y})\right], \tag{19}$$

where $\bar{y} = 1/T\sum_{i=1}^{T} y_i$ and $s_T = \sqrt{1/T\sum_{i=1}^{T}(y_i - \bar{y})^2}$. If y_t is iid with finite variance, then

$$\frac{1}{\sqrt{T}}Q_T \Rightarrow V,$$

where $\Rightarrow$ denotes weak convergence and V is the range of a Brownian bridge on the unit interval. Lo (1991) gives selected quantiles of V.

Lo (1991) pointed out that the R/S statistic is not robust to short range dependence. In particular, if y_t is autocorrelated (has short memory) then the limiting distribution of $Q_T/\sqrt{T}$ is V scaled by the square root of the long run variance of y_t. To allow for short range dependence in y_t, Lo (1991) modified the R/S statistic as follows

$$\tilde{Q}_T = \frac{1}{\hat{\sigma}_T(q)}\left[\max_{1\le k\le T}\sum_{j=1}^{k}(y_j - \bar{y}) - \min_{1\le k\le T}\sum_{j=1}^{k}(y_j - \bar{y})\right], \tag{20}$$

where the sample standard deviation is replaced by the square root of the Newey-West (Newey and West (1987)) estimate of the long run variance with bandwidth q.[10] Lo (1991) showed that if there is short memory but no long memory in y_t, $\tilde{Q}_T$ also converges to V, the range of a Brownian bridge. Breidt et al. (1998) found that (20) is effective for detecting long memory behavior in asset return volatility.

7.2 Two component GARCH model

In the covariance stationary GARCH model the conditional volatility will always mean revert to its long run level unconditional value. Recall the mean reverting form of the basic GARCH(1, 1) model in (11). In many empirical

[10] The long-run variance is the asymptotic variance of $\sqrt{T}(\bar{y} - \mu)$.

applications, the estimated mean reverting rate $\hat{a}_1 + \hat{b}_1$ is often very close to 1. For example, the estimated value of $a_1 + b_1$ from the GARCH(1,1) model for the S&P 500 index is 0.99 and the half life of a volatility shock implied by this mean reverting rate is $\ln(0.5)/\ln(0.956) = 76.5$ days. So the fitted GARCH(1,1) model implies that the conditional volatility is very persistent.

Engle and Lee (1999) suggested that the high persistence and long memory in volatility may be due to a time-varying long run volatility level. In particular, they suggested decomposing conditional variance into two components

$$\sigma_t^2 = q_t + s_t, \tag{21}$$

where q_t is a highly persistent long run component, and s_t is a transitory short run component. Long memory behavior can often be well approximated by a sum of two such components. A general form of the two components model that is based on a modified version of the PGARCH$(1, d, 1)$ is

$$\sigma_t^d = q_t^d + s_t^d, \tag{22}$$

$$q_t^d = \alpha_1|\epsilon_{t-1}|^d + \beta_1 q_{t-1}^d, \tag{23}$$

$$s_t^d = a_0 + \alpha_2|\epsilon_{t-1}|^d + \beta_2 s_{t-1}^d. \tag{24}$$

Here, the long run component q_t follows a highly persistent PGARCH$(1, d, 1)$ model and the transitory component s_t follows another PGARCH$(1, d, 1)$ model. For the two components to be separately identified the parameters should satisfy $1 < (\alpha_1 + \beta_1) < (\alpha_2 + \beta_2)$. It can be shown that the reduced form of the two components model is

$$\sigma_t^d = a_0 + (\alpha_1 + \alpha_2)|\epsilon_{t-1}|^d - (\alpha_1\beta_2 + \alpha_2\beta_1)|\epsilon_{t-2}|^d + (\beta_1 + \beta_2)\sigma_{t-1}^d - \beta_1\beta_2\sigma_{t-2}^d,$$

which is in the form of a constrained PGARCH$(2, d, 2)$ model. However, the two components model is not fully equivalent to the PGARCH$(2, d, 2)$ model because not all PGARCH$(2, d, 2)$ models have the component structure. Since the two components model is a constrained version of the PGARCH$(2, d, 2)$ model, the estimation of a two components model is often numerically more stable than the estimation of an unconstrained PGARCH$(2, d, 2)$ model.

7.3 Integrated GARCH model

The high persistence often observed in fitted GARCH(1,1) models suggests that volatility might be nonstationary implying that $a_1 + b_1 = 1$, in which case the GARCH(1,1) model becomes the integrated GARCH(1,1) or IGARCH(1,1) model. In the IGARCH(1,1) model the unconditional variance is not finite and so the model does not exhibit volatility mean rever-

Asset	$\tilde{Q}_T$	
	r_t^2	$\lvert r_t \rvert$
Microsoft	2.3916	3.4557
S&P 500	2.3982	5.1232

Table 9 Modified R/S Tests for Long Memory.

a_0	α_1	β_1	α_2	β_2	v	BIC
			Microsoft			
$2.86e^{-6}$	0.0182	0.9494	0.0985	0.7025		-20262
$[1.65e^{-6}]$	[0.0102]	[0.0188]	[0.0344]	[0.2017]		
$1.75e^{-6}$	0.0121	0.9624	0.0963	0.7416	6.924	-20501
$5.11e^{-7}$	[0.0039]	[0.0098]	[0.0172]	[0.0526]	[0.6975]	
			S&P 500			
$3.2e^{-8}$	0.0059	0.9848	0.1014	0.8076		−28113
$[1.14e^{-8}]$	[0.0013]	[0.0000]	[0.0221]	[0.0001]		
$1.06e^{-8}$	0.0055	0.9846	0.0599	0.8987	5.787	−28457
$[1.26e^{-8}]$	[0.0060]	[0.0106]	[0.0109]	[0.0375]	[0.5329]	

Notes: QML standard errors are in brackets.

Table 10 Estimates of Two Component GARCH(1,1) Models.

sion. However, it can be shown that the model is strictly stationary provided $E[\ln(a_1 z_t^2 + b_1)] < 0$. If the IGARCH(1,1) model is strictly stationary then the parameters of the model can still be consistently estimated by MLE.

Diebold and Lopez (1996) argued against the IGARCH specification for modeling highly persistent volatility processes for two reasons. First, they argue that the observed convergence toward normality of aggregated returns is inconsistent with the IGARCH model. Second, they argue that observed IGARCH behavior may result from misspecification of the conditional variance function. For example, a two components structure or ignored structural breaks in the unconditional variance (Lamoureux and Lastrapes (1990a) and Mikosch and Starica (2004)) can result in IGARCH behavior.

7.4 Long memory GARCH models for daily returns

Table 9 gives Lo's modified R/S statistic (20) applied to r_t^2 and $|r_t|$ for Microsoft and the S&P 500. The 1% right tailed critical value for the test is 2.098 (Lo (1991) Table 5.2) and so the modified R/S statistics are significant at the 1% level for both series providing evidence for long memory behavior in volatility.

Table 10 shows estimates of the two component GARCH(1,1) with $d = 2$, using Gaussian and Student's t errors, for the daily returns on Microsoft and the S&P 500. Notice that the BIC values are smaller than the BIC values for the unconstrained GARCH(2,2) models given in Table 3, which confirms the better numerical stability of the two component model. For both series,

the two components are present and satisfy $1 < (\alpha_1 + \beta_1) < (\alpha_2 + \beta_2)$. For Microsoft, the half-lives of the two components from the Gaussian (Student's t) models are 21 (26.8) days and 3.1 (3.9) days, respectively. For the S&P 500, the half-lives of the two components from the Gaussian (Student's t) models are 75 (69.9) days and 7.3 (16.4) days, respectively.

8 GARCH Model Prediction

An important task of modeling conditional volatility is to generate accurate forecasts for both the future value of a financial time series as well as its conditional volatility. Volatility forecasts are used for risk management, option pricing, portfolio allocation, trading strategies and model evaluation. Since the conditional mean of the general GARCH model (10) assumes a traditional ARMA form, forecasts of future values of the underlying time series can be obtained following the traditional approach for ARMA prediction. However, by also allowing for a time varying conditional variance, GARCH models can generate accurate forecasts of future volatility, especially over short horizons. This Section illustrates how to forecast volatility using GARCH models.

8.1 GARCH and forecasts for the conditional mean

Suppose one is interested in forecasting future values of y_T in the standard GARCH model described by (2), (3) and (6). For simplicity assume that $E_T[y_{T+1}] = c$. Then the minimum mean squared error $h-$ step ahead forecast of y_{T+h} is just c, which does not depend on the GARCH parameters, and the corresponding forecast error is

$$\epsilon_{T+h} = y_{T+h} - E_T[y_{T+h}].$$

The conditional variance of this forecast error is then

$$\text{var}_T(\epsilon_{T+h}) = E_T[\sigma^2_{T+h}],$$

which does depend on the GARCH parameters. Therefore, in order to produce confidence bands for the $h-$step ahead forecast the $h-$step ahead volatility forecast $E_T[\sigma^2_{T+h}]$ is needed.

8.2 Forecasts from the GARCH(1,1) model

For simplicity, consider the basic GARCH(1, 1) model (7) where $\epsilon_t = z_t\sigma_t$ such that $z_t \sim$ iid $(0, 1)$ and has a symmetric distribution. Assume the model is to be estimated over the time period $t = 1, 2, \cdots, T$. The optimal, in terms of mean-squared error, forecast of σ^2_{T+k} given information at time T is $E_T[\sigma^2_{T+k}]$ and can be computed using a simple recursion. For $k = 1$,

$$\begin{aligned} E_T[\sigma^2_{T+1}] &= a_0 + a_1 E_T[\epsilon^2_T] + b_1 E_T[\sigma^2_T] \\ &= a_0 + a_1\epsilon^2_T + b_1\sigma^2_T, \end{aligned} \tag{25}$$

where it assumed that ϵ^2_T and σ^2_T are known[11]. Similarly, for $k = 2$

$$\begin{aligned} E_T[\sigma^2_{T+2}] &= a_0 + a_1 E_T[\epsilon^2_{T+1}] + b_1 E_T[\sigma^2_{T+1}] \\ &= a_0 + (a_1 + b_1)E_T[\sigma^2_{T+1}]. \end{aligned}$$

since $E_T[\epsilon^2_{T+1}] = E_T[z^2_{T+1}\sigma^2_{T+1}] = E_T[\sigma^2_{T+1}]$. In general, for $k \geq 2$

$$\begin{aligned} E_T[\sigma^2_{T+k}] &= a_0 + (a_1 + b_1)E_T[\sigma^2_{T+k-1}] \\ &= a_0 \sum_{i=0}^{k-1}(a_1 + b_1)^i + (a_1 + b_1)^{k-1}(a_1\epsilon^2_T + b_1\sigma^2_T). \end{aligned} \tag{26}$$

An alternative representation of the forecasting equation (26) starts with the mean-adjusted form

$$\sigma^2_{T+1} - \bar{\sigma}^2 = a_1(\epsilon^2_T - \bar{\sigma}^2) + b_1(\sigma^2_T - \bar{\sigma}^2),$$

where $\bar{\sigma}^2 = a_0/(1 - a_1 - b_1)$ is the unconditional variance. Then by recursive substitution

$$E_T[\sigma^2_{T+k}] - \bar{\sigma}^2 = (a_1 + b_1)^{k-1}(E[\sigma^2_{T+1}] - \bar{\sigma}^2). \tag{27}$$

Notice that as $k \to \infty$, the volatility forecast in (26) approaches $\bar{\sigma}^2$ if the GARCH process is covariance stationary and the speed at which the forecasts approaches $\bar{\sigma}^2$ is captured by $a_1 + b_1$.

The forecasting algorithm (26) produces forecasts for the conditional variance σ^2_{T+k}. The forecast for the conditional volatility, σ_{T+k}, is usually defined as the square root of the forecast for σ^2_{T+k}.

The GARCH(1,1) forecasting algorithm (25) is closely related to an exponentially weighted moving average (EWMA) of past values of ϵ^2_t. This type of forecast is commonly used by RiskMetrics (Morgan (1997)). The EWMA forecast of σ^2_{T+1} has the form

[11] In practice, a_0, a_1, b_1, ϵ_T and σ^2_T are the fitted values computed from the estimated GARCH(1,1) model instead of the unobserved "true" values.

$$\sigma^2_{T+1,EWMA} = (1-\lambda)\sum_{s=0}^{\infty}\lambda^s\epsilon^2_{t-s} \tag{28}$$

for $\lambda \in (0,1)$. In (28), the weights sum to one, the first weight is $1-\lambda$, and the remaining weights decline exponentially. To relate the EWMA forecast to the GARCH(1,1) formula (25), (28) may be re-expressed as

$$\sigma^2_{T+1,EWMA} = (1-\lambda)\epsilon^2_T + \lambda\sigma^2_{T,EWMA} = \epsilon^2_T + \lambda(\sigma^2_{T,EWMA} - \epsilon^2_T),$$

which is of the form (25) with $a_0 = 0$, $a_1 = 1-\lambda$ and $b_1 = \lambda$. Therefore, the EWMA forecast is equivalent to the forecast from a restricted IGARCH(1,1) model. It follows that for any $h > 0$, $\sigma^2_{T+h,EWMA} = \sigma^2_{T,EWMA}$. As a result, unlike the GARCH(1,1) forecast, the EWMA forecast does not exhibit mean reversion to a long-run unconditional variance.

8.3 Forecasts from asymmetric GARCH(1,1) models

To illustrate the asymmetric effects of leverage on forecasting, consider (cf. (17)) the TGARCH(1,1) at time T

$$\sigma^2_T = a_0 + a_1\epsilon^2_{T-1} + \gamma_1 S_{T-1}\epsilon^2_{T-1} + b_1\sigma^2_{T-1}.$$

Assume that ϵ_t has a symmetric distribution about zero. The forecast for $T+1$ based on information at time T is

$$E_T[\sigma^2_{T+1}] = a_0 + a_1\epsilon^2_T + \gamma_1 S_T\epsilon^2_T + b_1\sigma^2_T,$$

where it assumed that ϵ^2_T, S_T and σ^2_T are known. Hence, the TGARCH(1,1) forecast for $T+1$ will be different than the GARCH(1,1) forecast if $S_T = 1$ ($\epsilon_T < 0$). The forecast at $T+2$ is

$$\begin{aligned} E_T[\sigma^2_{T+2}] &= a_0 + a_1 E_T[\epsilon^2_{T+1}] + \gamma_1 E_T[S_{T+1}\epsilon^2_{T+1}] + b_1 E_T[\sigma^2_{T+1}] \\ &= a_0 + \left(\frac{\gamma_1}{2} + a_1 + b_1\right) E_T[\sigma^2_{T+1}], \end{aligned}$$

which follows since $E_T[S_{T+1}\epsilon^2_{T+1}] = E_T[S_{T+1}]E_T[\epsilon^2_{T+1}] = \frac{1}{2}E_T[\sigma^2_{T+1}]$. Notice that the asymmetric impact of leverage is present even if $S_T = 0$. By recursive substitution for the forecast at $T+h$ is

$$E_T[\sigma^2_{T+h}] = a_0 + \left(\frac{\gamma_1}{2} + a_1 + b_1\right)^{h-1} E_T[\sigma^2_{T+1}], \tag{29}$$

which is similar to the GARCH(1,1) forecast (26). The mean reverting form (29) is

$$E_T[\sigma^2_{T+h}] - \bar{\sigma}^2 = \left(\frac{\gamma_1}{2} + a_1 + b_1\right)^{h-1} \left(E_T[\sigma^2_{T+h}] - \bar{\sigma}^2\right)$$

where $\bar{\sigma}^2 = a_0/(1 - \frac{\gamma_1}{2} - a_1 - b_1)$ is the long run variance.

Forecasting algorithms for σ^d_{T+h} in the PGARCH$(1, d, 1)$ and for $\ln \sigma^2_{T+h}$ in the EGARCH(1,1) follow in a similar manner and the reader is referred to Ding et al. (1993), and Nelson (1991) for further details.

8.4 Simulation-based forecasts

The forecasted volatility can be used together with forecasted series values to generate confidence intervals of the forecasted series values. In many cases, the forecasted volatility is of central interest, and confidence intervals for the forecasted volatility can be obtained as well. However, analytic formulas for confidence intervals of forecasted volatility are only known for some special cases (see Baillie and Bollerslev (1992)). In models for which analytic formulas for confidence intervals are not known, a simulation-based method can be used to obtain confidence intervals for forecasted volatility from any GARCH that can be simulated. To obtain volatility forecasts from a fitted GARCH model, simply simulate σ^2_{T+k} from the last observation of the fitted model. This process can be repeated many times to obtain an "ensemble" of volatility forecasts. The point forecast of σ^2_{T+k} may then be computed by averaging over the simulations, and a 95% confidence interval may be computed using the 2.5% and 97.5% quantiles of the simulation distribution, respectively.

8.5 Forecasting the volatility of multiperiod returns

In many situations, a GARCH model is fit to daily continuously compounded returns $r_t = \ln(P_t) - \ln(P_{t-1})$, where P_t denotes the closing price on day t. The resulting GARCH forecasts are for daily volatility at different horizons. For risk management and option pricing with stochastic volatility, volatility forecasts are needed for multiperiod returns. With continuously compounded returns, the $h-$day return between days T and $T + h$ is simply the sum of h single day returns

$$r_{T+h}(h) = \sum_{j=1}^{h} r_{T+j}.$$

Assuming returns are uncorrelated, the conditional variance of the $h-$period return is then

$$\text{var}_T(r_{T+h}(h)) = \sigma_T^2(h) = \sum_{j=1}^{h} \text{var}_T(r_{T+j}) = E_T[\sigma_{T+1}^2] + \cdots + E_T[\sigma_{T+h}^2]. \tag{30}$$

If returns have constant variance $\bar{\sigma}^2$, then $\sigma_T^2(h) = h\bar{\sigma}^2$ and $\sigma_T(h) = \sqrt{h}\bar{\sigma}$. This is known as the "square root of time" rule as the $h-$day volatility scales with $\sqrt{h}$. In this case, the $h-$day variance per day, $\sigma_T^2(h)/h$, is constant. If returns are described by a GARCH model then the square root of time rule does not necessarily apply. To see this, suppose returns follow a GARCH(1,1) model. Plugging the GARCH(1,1) model forecasts (27) for $E_T[\sigma_{T+1}^2], \ldots, E_T[\sigma_{T+h}^2]$ into (30) gives

$$\sigma_T^2(h) = h\bar{\sigma}^2 + (E[\sigma_{T+1}^2] - \bar{\sigma}^2)\left[\frac{1-(a_1+b_1)^h}{1-(a_1+b_1)}\right]$$

For the GARCH(1,1) process the square root of time rule only holds if $E[\sigma_{T+1}^2] = \bar{\sigma}^2$. Whether $\sigma_T^2(h)$ is larger or smaller than $h\bar{\sigma}^2$ depends on whether $E[\sigma_{T+1}^2]$ is larger or smaller than $\bar{\sigma}^2$.

8.6 Evaluating volatility predictions

GARCH models are often judged by their out-of-sample forecasting ability, see Clements (2005) for an overview. This forecasting ability can be measured using traditional forecast error metrics as well as with specific economic considerations such as value-at-risk violations, option pricing accuracy, or portfolio performance. Out-of-sample forecasts for use in model comparison are typically computed using one of two methods. The first method produces recursive forecasts. An initial sample using data from $t = 1, \ldots, T$ is used to estimate the models, and $h-$step ahead out-of-sample forecasts are produced starting at time T. Then the sample is increased by one, the models are re-estimated, and $h-$step ahead forecasts are produced starting at $T + 1$. This process is repeated until no more $h-$step ahead forecasts can be computed. The second method produces rolling forecasts. An initial sample using data from $t = 1, \ldots, T$ is used to determine a window width T, to estimate the models, and to form $h-$step ahead out-of-sample forecasts starting at time T. Then the window is moved ahead one time period, the models are re-estimated using data from $t = 2, \ldots, T + 1$, and $h-$step ahead out-of-sample forecasts are produced starting at time $T + 1$. This process is repeated until no more $h-$step ahead forecasts can be computed.

8.6.1 Traditional forecast evaluation statistics

Let $E_{i,T}[\sigma^2_{T+h}]$ denote the $h-$step ahead forecast of σ^2_{T+h} at time T from GARCH model i using either recursive or rolling methods. Define the corresponding forecast error as $e_{i,T+h|T} = E_{i,T}[\sigma^2_{T+h}] - \sigma^2_{T+h}$. Common forecast evaluation statistics based on N out-of-sample forecasts from $T = T+1,\ldots,T+N$ are

$$\text{MSE}_i = \frac{1}{N}\sum_{j=T+1}^{T+N} e^2_{i,j+h|j},$$

$$\text{MAE}_i = \frac{1}{N}\sum_{j=T+1}^{T+N} \left|e_{i,j+h|j}\right|,$$

$$\text{MAPE}_i = \frac{1}{N}\sum_{j=T+1}^{T+N} \frac{\left|e_{i,j+h|j}\right|}{\sigma_{j+h}}.$$

The model which produces the smallest values of the forecast evaluation statistics is judged to be the best model. Of course, the forecast evaluation statistics are random variables and a formal statistical procedure should be used to determine if one model exhibits superior predictive performance.

Diebold and Mariano (1995) proposed a simple procedure to test the null hypothesis that one model has superior predictive performance over another model based on traditional forecast evaluation statistics. Let $\{e_{1,j+h|j}\}_{T+1}^{T+N}$, and $\{e_{2,j+h|j}\}_{T+1}^{T+N}$ denote forecast errors from two different GARCH models. The accuracy of each forecast is measured by a particular loss function $L(e_{i,T+h|T})$, $i = 1,2$. Common choices are the squared error loss function $L(e_{i,T+h|T}) = \left(e_{i,T+h|T}\right)^2$ and the absolute error loss function $L(e_{i,T+h|T}) = \left|e_{i,T+h|T}\right|$. The Diebold-Mariano (DM) test is based on the loss differential

$$d_{T+h} = L(e_{1,T+h|T}) - L(e_{2,T+h|T}).$$

The null of equal predictive accuracy is $H_0 : E[d_{T+h}] = 0$.The DM test statistic is

$$\text{S} = \frac{\bar{d}}{\left(\widehat{\text{avar}}(\bar{d})\right)^{1/2}}, \tag{31}$$

where $\bar{d} = N^{-1}\sum_{j=T+1}^{T+N} d_{j+h}$, and $\widehat{\text{avar}}(\bar{d})$ is a consistent estimate of the asymptotic variance of $\sqrt{N}\bar{d}$. Diebold and Mariano (1995) recommend using the Newey-West estimate for $\widehat{\text{avar}}(\bar{d})$ because the sample of loss differentials $\{d_{j+h}\}_{T+1}^{T+N}$ are serially correlated for $h > 1$. Under the null of equal predictive accuracy, S has an asymptotic standard normal distribution. Hence, the DM statistic can be used to test if a given forecast evaluation statistic (e.g. MSE_1)

for one model is statistically different from the forecast evaluation statistic for another model (e.g. MSE_2).

Forecasts are also often judged using the forecasting regression

$$\sigma^2_{T+h} = \alpha + \beta E_{i,T}[\sigma^2_{T+h}] + e_{i,T+h}. \tag{32}$$

Unbiased forecasts have $\alpha = 0$ and $\beta = 1$, and accurate forecasts have high regression R^2 values. In practice, the forecasting regression suffers from an errors-in-variables problem when estimated GARCH parameters are used to form $E_{i,T}[\sigma^2_{T+h}]$ and this creates a downward bias in the estimate of β. As a result, attention is more often focused on the R^2 from (32).

An important practical problem with applying forecast evaluations to volatility models is that the $h-$step ahead volatility σ^2_{T+h} is not directly observable. Typically, ϵ^2_{T+h} (or just the squared return) is used to proxy σ^2_{T+h} since $E_T[\epsilon^2_{T+h}] = E_T[z^2_{T+h}\sigma^2_{T+h}] = E_T[\sigma^2_{T+h}]$. However, ϵ^2_{T+h} is a very noisy proxy for σ^2_{T+h} since $\text{var}(\epsilon^2_{T+h}) = E[\sigma^4_{T+h}](\kappa - 1)$, where κ is the fourth moment of z_t, and this causes problems for the interpretation of the forecast evaluation metrics.

Many empirical papers have evaluated the forecasting accuracy of competing GARCH models using ϵ^2_{T+h} as a proxy for σ^2_{T+h}. Poon (2005) gave a comprehensive survey. The typical findings are that the forecasting evaluation statistics tend to be large, the forecasting regressions tend to be slightly biased, and the regression R^2 values tend to be very low (typically below 0.1). In general, asymmetric GARCH models tend to have the lowest forecast evaluation statistics. The overall conclusion, however, is that GARCH models do not forecast very well.

Andersen and Bollerslev (1998) provided an explanation for the apparent poor forecasting performance of GARCH models when ϵ^2_{T+h} is used as a proxy for σ^2_{T+h} in (32). For the GARCH(1,1) model in which z_t has finite kurtosis κ, they showed that the population R^2 value in (32) with $h = 1$ is equal to

$$R^2 = \frac{a_1^2}{1 - b_1^2 - 2a_1 b_1},$$

and is bounded from above by $1/\kappa$. Assuming $z_t \sim N(0,1)$, this upper bound is 1/3. With a fat-tailed distribution for z_t the upper bound is smaller. Hence, very low R^2 values are to be expected even if the true model is a GARCH(1,1). Moreover, Hansen and Lunde (2004) found that the substitution of ϵ^2_{T+h} for σ^2_{T+h} in the evaluation of GARCH models using the DM statistic (31) can result in inferior models being chosen as the best with probability one. These results indicate that extreme care must be used when interpreting forecast evaluation statistics and tests based on ϵ^2_{T+h}. If high frequency intraday data are available, then instead of using ϵ^2_{T+h} to proxy σ^2_{T+h} Andersen and Bollerslev (1998) suggested using the so-called realized variance

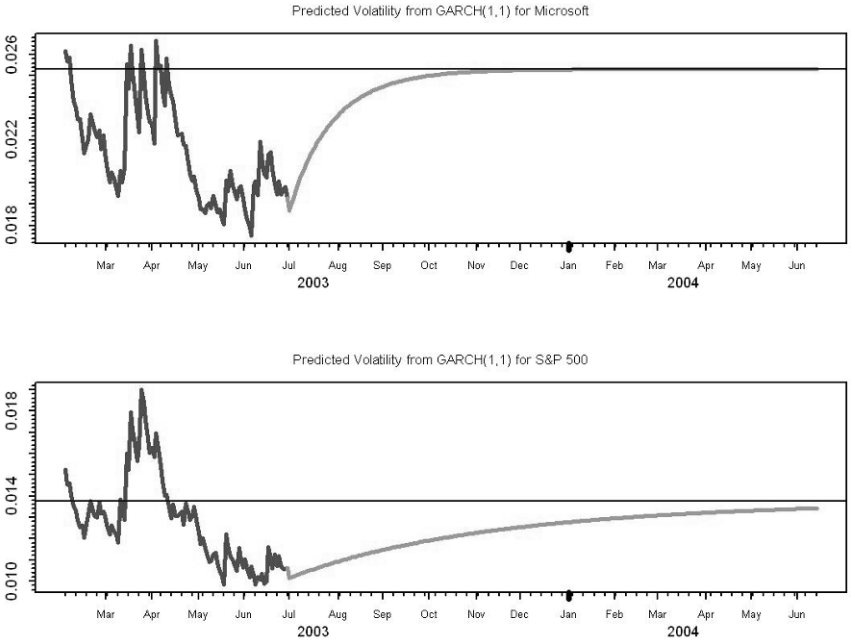

Fig. 8 Predicted Volatility from GARCH(1,1) Models

$$RV_{t+h}^{m} = \sum_{j=1}^{m} r_{t+h,j}^{2},$$

where $\{r_{T+h,1}, \ldots, r_{T+h,m}\}$ denote the squared intraday returns at sampling frequency $1/m$ for day $T + h$. For example, if prices are sampled every 5 minutes and trading takes place 24 hours per day then there are $m = 288$ 5-minute intervals per trading day. Under certain conditions (see Andersen et al. (2003)), RV_{t+h}^{m} is a consistent estimate of σ_{T+h}^{2} as $m \to \infty$. As a result, RV_{t+h}^{m} is a much less noisy estimate of σ_{T+h}^{2} than ϵ_{T+h}^{2} and so forecast evaluations based on RV_{t+h}^{m} are expected to be much more accurate than those based on ϵ_{T+h}^{2}. For example, in evaluating GARCH(1,1) forecasts for the Deutschemark-US daily exchange rate, Andersen and Bollerslev (1998) reported R^2 values from (32) of $0.047, 0.331$ and 0.479 using ϵ_{T+1}^{2}, RV_{T+1}^{24} and RV_{T+1}^{288}, respectively.

8.7 Forecasting the volatility of Microsoft and the S&P 500

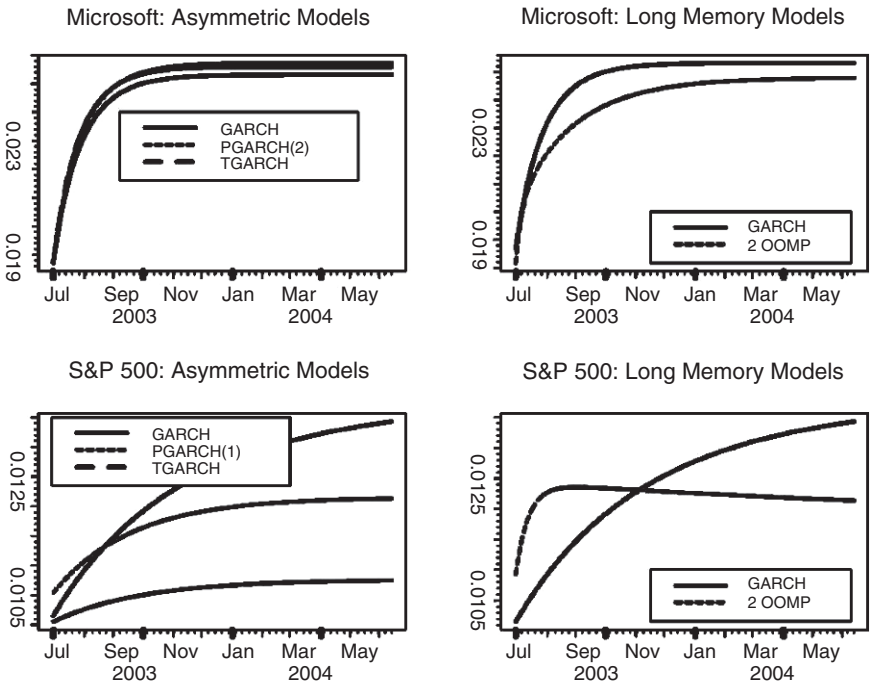

Fig. 9 Predicted Volatility from Competing GARCH Models.

Figure 8 shows $h-$day ahead volatility predictions ($h = 1, \ldots, 250$) from the fitted GARCH(1,1) models with normal errors for the daily returns on Microsoft and the S&P 500. The horizontal line in the figures represents the estimated unconditional standard deviation from the fitted models. At the beginning of the forecast period, $\hat{\sigma}_T < \widehat{\bar{\sigma}}$ for both series and so the forecasts revert upward toward the unconditional volatility. The speed of volatility mean reversion is clearly shown by the forecast profiles. The forecasts for Microsoft revert to the unconditional level after about four months, whereas the forecasts for the S&P 500 take over one year.

Figure 9 shows the volatility forecasts from the asymmetric and long memory GARCH(1,1) models, and Table 11 gives the unconditional volatility from the estimated models. For Microsoft, the forecasts and unconditional volatilities from the different models are similar. For the S&P 500, in contrast, the

	Error pdf	GARCH	TGARCH	PGARCH
MSFT	Gaussian	0.0253	0.0257	0.0256
	Student's t	0.0247	0.0253	0.0250
S&P 500	Gaussian	0.0138	0.0122	0.0108
	Student's t	0.0138	0.0128	0.0111

Table 11 Unconditional Volatilities from Estimated GARCH(1,1) Models.

forecasts and unconditional volatilities differ considerably across the models.

9 Final Remarks

This chapter surveyed some of the practical issues associated with estimating univariate GARCH models and forecasting volatility. Some practical issues associated with the estimation of multivariate GARCH models and forecasting of conditional covariances are given in Silvennoinen and Teräsvirta (2008).

References

Alexander, C. (2001): *Market Models: A Guide to Financial Data Analysis*. John Wiley & Sons, Chichester.

Andersen, T. and Bollerslev, T. (1998): Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. *International Economic Review* **39**, 885–905.

Andersen, T., Bollerslev, T., Christoffersen, P.K. and Diebold, F.X. (2006): Volatility Forecasting. *In: Elliott, G., Granger, C.W.J. and Timmermann, A. (Eds.): Handbook of Economic Forecasting*. North Holland, Amsterdam.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2003): The Distribution of Exchange Rate Volatility. *Journal of the American Statistical Association* **96**, 42–55.

Andreou, E. and Ghysels, E. (2002): Rolling Volatility Estimators: Some new Theoretical, Simulation and Empirical Results. *Journal of Business and Economic Statistics* **20**, 363–376.

Baillie, R.T. and Bollerslev, T. (1992): Prediction in Dynamic Models with Time Dependent Conditional Variances. *Journal of Econometrics* **52**, 91–113.

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996): Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **74**, 3–30.

Beran, J. (1994): *Statistics for Long Memory Processes*. Chapman and Hall, New York.

Bera, A.K. and Higgins, M.L. (1995): On ARCH Models: Properties, Estimation and Testing. *Journal of Economic Surveys* **7**, 305–362.

Black, F. (1976): Studies in Stock Price Volatility Changes. *Proceedings of the 1976 Business Meeting of the Business and Economics Statistics Section*, American Statistical Association, 177–181.

Blair, B., Poon, S.-H., and Taylor, S.J. (2001): Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and High Frequency Index Returns. *Journal of Econometrics* **105**, 5–26.

Bollerslev, T. (1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T. (1987): A Conditional Heteroskedastic Time Series Model for Speculative Prices Rates of Return. *Review of Economics and Statistics* **69**, 542–547.

Bollerslev, T., Engle, R.F. and Nelson, D.B. (1994): ARCH Models. *In: Engle, R.F. and McFadden, D.L. (Ed.): Handbook of Econometrics* **4**. Elsevier Science B. V., Amsterdam.

Bollerslev, T. and Ghysels, E. (1996): Periodic Autoregressive Conditional Heteroscedasticity. *Journal of Business and Economic Statistics* **14**, 139–151.

Bollerslev, T. and Wooldridge, T. (1992): Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-varying Covariances. *Econometric Reviews* **11**, 143–172.

Bomfim, A.N. (2003): Pre-Announcement Effects, News Effects, and Volatility: Monetary Policy and the Stock Market. *Journal of Banking and Finance* **27**, 133–151.

Breidt, F.J., Crato, N. and de Lima, P. (1998): The Detection and Estimation of Long Memory in Stochastic Volatility. *Journal of Econometrics* **73**, 325–348.

Brooks, C. (1997): GARCH Modeling in Finance: A Review of the Software Options. *Economic Journal* **107**, 1271–1276.

Brooks, C., Burke, S. and Pesand, G. (2001): Benchmarks and the Accuracy of GARCH Model Estimation. *International Journal of Forecasting* **17**, 45–56.

Chen, F., Yu, W.-C. and Zivot, E. (2008): Predicting Stock Volatility Using After-Hours Information. *Unpublished manuscript, Department of Economics, University of Washington.*

Clements, M.P. (2005): Evaluating Econometric Forecasts of Economic and Financial Variables. *Palgrave Texts in Econometrics, Palgrave Macmillan, Houndmills, UK.*

Conrad, C. and Haag, B.R. (2006): Inequality Constraints in the Fractionally Integrated GARCH Model. *Journal of Financial Econometrics* **4**, 413–449.

Davis, R. and T. Mikosch (2008): Extreme value theory for GARCH processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 186–200. Springer, New York.

de Lima, P.J.F. (1996): Nuisance Parameter Free Properties of Correlation Integral Based Statistics. *Econometric Reviews* **15**, 237–259.

Diebold, F.X. (1988): *Empirical Modeling of Exchange Rate Behavior.* Springer, New York.

Diebold, F.X. and Lopez, J.A. (1996): Modeling Volatility Dynamics. *In: Hoover, K. (Ed.): Macroeconomics: Developments, Tensions and Prospects.* Kluwer, Boston.

Diebold, F.X. and Mariano, R.S. (1995): Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* **13**, 253–263.

Diebold, F.X. and Schuermann, T. (1993): Exact Maximum Likelihood Estimation of ARCH Models. *Unpublished manuscript, Department of Economics, University of Pennsylvania.*

Ding, Z. and Granger C.W.J. (1996): Modeling Volatility Persistence of Speculative Returns: A New Approach. *Journal of Econometrics* **73**, 185–215.

Ding, Z., Granger, C.W.J. and Engle, R.F. (1993): A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance* **1**, 83–106.

Drost, F.C. and Nijman, T.E. (1993): Temporal Aggregation of GARCH Processes. *Econometrica* **61**, 909–927.

Engle, R.F. (1982): Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **50**, 987–1007.

Engle, R.F. (2001): GARCH 101: The Use of ARCH/GARCH Models in Applied Economics. *Journal of Economic Perspectives* **15**, 157–168.

Engle, R.F. and González-Rivera, G. (1991): Semiparametric ARCH models. *Journal of Business and Economic Statistics* **9**, 345–60.

Engle, R.F. and Kroner, K. (1995): Multivariate Simultaneous Generalised ARCH. *Econometric Theory* **11**, 122–150.

Engle, R.F. and Lee, G.J. (1999): A Long-Run and Short-Run Component Model of Stock Return Volatility. *In: Engle, R.F. and White, H. (Eds.): Cointegration, Causality, and Forecasting*. Oxford University Press, Oxford.

Engle, R.F. and Mezrich, J. (1995): Grappling with GARCH. *RISK* **8**, 112–117.

Engle, R.F. and Ng, V. (1993): Measuring and Testing the Impact of News on Volatility. *Journal of Finance* **48**, 1749–78.

Engle, R.F. and Patton, A. (2001): What Good is a Volatility Model? *Quantitative Finance* **1**, 237–245.

Fiorentini, G., Calzolari, G. and Panattoni, L. (1996): Analytic Derivatives and the Computation of GARCH Estimates. *Journal of Applied Econometrics* **11**, 399–417.

Fernandez, C. and Steel, M. (1998): On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association* **93**, 359–371.

Flannery, M. and Protopapadakis, A. (2002): Macroeconomic Factors Do Influence Aggregate Stock Returns. *The Review of Financial Studies* **15**, 751–782.

Francq, C. and Zakoïan, J.-M. (2008): A tour in the asymptotic theory of GARCH estimation. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 85–111. Springer, New York.

Franses, P.H. and van Dijk, D. (2000): *Non-Linear Time Series Models in Empirical Finance*. Cambridge University Press, Cambridge.

Gallant, A.R. and Tauchen, G. (2001): SNP: A Program for Nonparametric Time Series Analysis, Version 8.8, User's Guide. *Unpublished manuscrip, University of North Carolina at Chapel Hill*.

Gallo, G.M. and Pacini, B. (1997): Early News is Good News: The Effect of Market Opening on Market Volatility. *Studies in Nonlinear Dynamics and Econometrics* **2**, 115–131.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993): On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance* **48**, 1779–1801.

Hamilton, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Hansen, P. and Lunde, A. (2004): A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1) Model? *Journal of Applied Econometrics* **20**, 873–889.

Hansen, P. and Lunde, A. (2006): Consistent Ranking of Volatility Models. *Journal of Econometrics* **131**, 97–121.

He, C. and Teräsvirta, T. (1999a): Properties of Moments of a Family of GARCH Processes. *Journal of Econometrics* **92**, 173–192.

He, C. and Teräsvirta, T. (1999b): Fourth Moment Structure of the GARCH(p, q) Process. *Econometric Theory* **15**, 824–846.

Hurst, H.E. (1951): Long Term Storage Capacity of Reservoirs. *Transactions of the American Society of Civil Engineers* **116**, 770–799.

Morgan, J.P. (1997): *RiskMetrics, Technical Documents, 4th Edition*. New York.

Jensen, T. and Rahbek, A. (2004): Asymptotic Normality of the QML Estimator of ARCH in the Nonstationary Case. *Econometrica* **72**, 641–646.

Kristensen, D. and Rahbek, A. (2005): Asymptotics of the QMLE for a Class of ARCH(q) Models. *Econometric Theory* **21**, 946–961.

Laurent, S. and Peters, J.-P. (2002): G@RCH 2.0: An Ox Package for Estimating and Forecasting Various ARCH Models. *Journal of Economic Surveys* **16**, 447–485.

Lamoureux, C.G. and Lastrapes, W.D. (1990a): Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects. *The Journal of Finance* **45**, 221–229.

Lamoureux, C.G. and Lastrapes, W.D. (1990b): Persistence in Variance, Structural Change and the GARCH Model. *Journal of Business and Economic Statistics* **8**, 225–234.

Lee, J.H.H. and Hansen, B.E. (1993): Asymptotic Theory for the GARCH(1,1) Quasi-Maximum Likelihood Estimator. *Econometric Theory* **10**, 29–52.

Lee, J.H.H. and King, M.L. (1993): A Locally Most Powerful Based Score Test for ARCH and GARCH Regression Disturbances. *Journal of Business and Economic Statistics* **7**, 259–279.

Leeb, H. and Pötscher, B.M. (2008): Model selection. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 888–925. Springer, New York.

Lo, A.W. (1991): Long Term Memory in Stock Market Prices. *Econometrica* **59**, 1279–1313.

Lumsdaine, R.L. (1992): Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1,1) and Covariance Stationary GARCH(1,1) Models. *Econometrica* **64**, 575–596.

Lumsdaine, R.L. and Ng, S. (1999): Testing for ARCH in the Presence of a Possibly Misspecified Conditional Mean. *Journal of Econometrics* **93**, 257–279.

Lundbergh, S. and Teräsvirta, T. (2002): Evaluating GARCH Models. *Journal of Econometrics* **110**, 417–435.

Ma, J., Nelson, C.R. and Startz, R. (2007): Spurious Inference in the GARCH(1,1) Model When It Is Weakly Identified. *Studies in Nonlinear Dynamics and Econometrics* **11**, Article 1.

Mandelbrot, B.B. (1975): Limit Theorems on the Self-Normalized Range for Weakly and Strongly Dependent Processes. *Zeitschrift für Wahr- scheinlichkeitstheorie und verwandte Gebiete* **31**, 271–285.

Martens, M. (2002): Measuring and Forecasting S&P 500 Index-Futures Volatility Using High Frequency Data. *The Journal of Futures Markets* **22**, 497–518.

McCullough, B.D. and Renfro, C.G. (1999): Benchmarks and Software Standards: A Case Study of GARCH Procedures. *Journal of Economic and Social Measurement* **25**, 59–71.

Mikosch, T. and Starica, C. (2004): Non-stationarities in Financial Time Series, the Long-Range Dependence and the IGARCH Effects. *Review of Economics and Statistics* **86**, 378–384.

Nelson, D. B. (1991): Conditional Heteroskedasticity in Asset Returns: a New Approach. *Econometrica* **59**, 347–370.

Nelson, D. B. and Cao, C. Q. (1992): Inequality Constraints in the Univariate GARCH Model. *Journal of Business and Economic Statistics* **10**, 229–235.

Newey, W.K. and West, K.D. (1987): A Simple Positive Semidefinite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **55**, 703–708.

Pagan, A. (1996): The Econometrics of Financial Markets. *Journal of Empirical Finance* **3**, 15–102.

Pagan, A. and Schwert G.W. (1990): Alternative Models for Conditional Volatility. *Journal of Econometrics* **45**, 267–290.

Palm, F.C. (1996): GARCH models of Volatility. *In: Maddala, G.S. and Rao, C.R. (Eds.): Handbook of Statistics. Statistical Methods in Finance* **14**, 209–240. North Holland.

Poon, S.-H. (2005): *A Practical Guide to Forecasting Financial Market Volatility* John Wiley & Sons, New York.

Poon, S.-H. and Granger, C. (2003): Forecasting Financial Market Volatility: A Review. *Journal of Economic Literature* **41**, 478–539.

Poon, S.-H. and Granger, C. (2005): Practical Issues in Forecasting Volatility. *Financial Analysts Journal* **61**, 45–56.

Silvennoinen, A. and Teräsvirta, T. (2008): Multivariate GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 201–229 Springer, New York.

Straumann, D. (2005): Estimation in Conditionally Heteroskedastic Time Series Models. *Lecture Notes in Statistics* **181**, Springer, Berlin.

Taylor, S.J. and Xu, X. (1997): The Incremental Volatilty Information in One Million Foreign Exchange Quotations. *Journal of Empirical Finance* **4**, 317–340.

Teräsvirta, T. (2008): An introduction to univariate GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 17–42. Springer, New York.

Tsay, R.S. (2001): *Analysis of Financial Time Series* John Wiley & Sons, New York.

Venter, J.H. and de Jongh, P.J. (2002): Risk Estimation Using the Normal Inverse Gausian Distribution. *Journal of Risk* **4**, 1–23.

Weiss, A.A. (1986): Asymptotic Theory for ARCH Models: Estimation and Testing. *Econometric Theory* **2**, 107–131.

Zakoïan, M. (1994): Threshold Heteroscedastic Models. *Journal of Economic Dynamics and Control* **18**, 931–955.

Zivot, E. and Wang, J. (2005): *Modeling Financial Time Series with S-PLUS, Second Edition* Springer, New York.

Semiparametric and Nonparametric ARCH Modeling

Oliver B. Linton*

Abstract This paper surveys nonparametric approaches to modelling discrete time volatility. We cover functional form, error shape, memory, and relationship between mean and variance.

1 Introduction

The key properties of financial time series appear to be: (a) Marginal distributions have heavy tails and thin centres (Leptokurtosis); (b) the scale or spread appears to change over time; (c) Return series appear to be almost uncorrelated over time but to be dependent through higher moments. See Mandelbrot (1963) and Fama (1965) for some early discussions. The traditional linear models like the autoregressive moving average class do not capture all these phenomena well. This is the motivation for using nonlinear models. This chapter is about the nonparametric approach.

2 The GARCH Model

Stochastic volatility models are of considerable current interest in empirical finance following the seminal work of Engle (1982). Perhaps still the most

Oliver B. Linton
Department of Economics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom, http://econ.lse.ac.uk/staff/olinton/ e-mail: o.linton@lse.ac.uk

* Thanks to the ESRC for financial support.

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_6,

popular version is Bollerslev's (1986) GARCH(1,1) model in which the conditional variance σ_t^2 of a martingale difference sequence y_t is

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2, \tag{1}$$

where the ARCH(1) process corresponds to $\beta = 0$. This model has been extensively studied and generalized in various ways, see the review of Bollerslev, Engle, and Nelson (1994). Following Drost and Nijman (1993), we can give three interpretations to (1). The *strong* form GARCH(1,1) process arises when

$$\frac{y_t}{\sigma_t} = \varepsilon_t \tag{2}$$

is i.i.d. with mean zero and variance one, where σ_t^2 is defined in (1). The most common special case is where ε_t are also standard normal. The *semi-strong* form arises when for ε_t in (2)

$$E(\varepsilon_t \,|\mathcal{F}_{t-1}) = 0 \text{ and } E(\varepsilon_t^2 - 1 \,|\mathcal{F}_{t-1}) = 0, \tag{3}$$

where $\mathcal{F}_{t-1}$ is the sigma field generated by the entire past history of the y process. Finally, there is a *weak* form in which σ_t^2 is defined as a projection on a certain subspace, so that the actual conditional variance may not coincide with (1). The properties of the strong GARCH process are well understood, and under restrictions on the parameters $\theta = (\omega, \beta, \gamma)$ it can be shown to be strictly positive with probability one, to be weakly and/or strictly stationary, and to be geometrically mixing and ergodic. The weaknesses of the model are by now well documented.

3 The Nonparametric Approach

There are several different ways in which nonparametric components have been introduced into stochastic volatility models. This work was designed to overcome some of the restrictiveness of the parametric assumptions in Gaussian strong GARCH models.

3.1 Error density

Estimation of the strong GARCH process usually proceeds by specifying that the error density ε_t is standard normal and then maximizing the (conditional on initial values) Gaussian likelihood function. It has been shown that the resulting estimators are consistent and asymptotically normal under a variety of conditions. It is not required that the error terms actually be normal or even i.i.d., but if this condition does not hold the resulting estimator is not

efficient. In many cases, there is evidence that the standardized residuals from estimated GARCH models are not normally distributed, especially for high frequency financial time series. Engle and Gonzalez-Rivera (1991) initiated the study of semiparametric models in which ε_t is i.i.d. with some density f that may be non-normal, thus suppose that

$$y_t = \varepsilon_t \sigma_t$$
$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma y_{t-1}^2,$$

where ε_t is i.i.d. with density f of unknown functional form. There is evidence that the density of the standardized residuals $\varepsilon_t = y_t/\sigma_t$ is non-Gaussian. One can obtain more efficient estimates of the parameters of interest by estimating f nonparametrically. Linton (1993) and Drost and Klaassen (1997) develop kernel based estimates and establish the semiparametric efficiency bounds for estimation of the parameters. In some cases, e.g., if f is symmetric about zero, it is possible to adaptively estimate some parameters, i.e., one can achieve the same asymptotic efficiency as if one knew the error density. Hafner and Rombouts (2006) have recently treated a number of multivariate cases and show that it is not generally possible to adapt, although one can achieve a semiparametric efficiency bound. These methods can often deliver efficiency gains but may not be robust to say dependent or time varying ε_t. In practice, the estimated density is quite heavy tailed but close to symmetric for stock returns. These semiparametric models can readily be applied to deliver value at risk and conditional value at risk measures based on the estimated density.

3.2 Functional form of volatility function

Another line of work has been to question the specific functional form of the volatility function, since estimation is not robust with respect to its specification. The news impact curve is the relationship between σ_t^2 and $y_{t-1} = y$ holding past values σ_{t-1}^2 constant at some level σ^2. This is an important relationship that describes how new information affects volatility. For the GARCH process, the news impact curve is

$$m(y, \sigma^2) = \omega + \gamma y^2 + \beta \sigma^2. \tag{4}$$

It is separable in σ^2, i.e., $\partial m(y, \sigma^2)/\partial \sigma^2$ does not depend on y, it is an even function of news y, i.e., $m(y, \sigma^2) = m(-y, \sigma^2)$, and it is a quadratic function of y with minimum at zero. The evenness property implies that $\mathrm{cov}(y_t^2, y_{t-j}) = 0$ for ε_t with distribution symmetric about zero.

Because of limited liability, we might expect that negative and positive shocks have different effects on the volatility of stock returns, for example.

The evenness of the GARCH process news impact curve rules out such 'leverage effects'. Nelson (1991) introduced the Exponential GARCH model to address this issue. Let $h_t = \log \sigma_t^2$ and let $h_t = \omega + \gamma \left[\theta \varepsilon_{t-1} + \delta \left|\varepsilon_{t-1}\right|\right] + \beta h_{t-1}$, where $\varepsilon_t = y_t/\sigma_t$ is i.i.d. with mean zero and variance one. This allows asymmetric effect of past shocks ε_{t-j} on current volatility, i.e., the news impact curve is allowed to be asymmetric. For example, $\mathrm{cov}(y_t^2, y_{t-j}) \neq 0$ even when ε_t is symmetric about zero. An alternative approach to allowing asymmetric news impact curve is the Glosten, Jeganathan and Runkle (1994) model $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 + \delta y_{t-1}^2 1(y_{t-1} < 0)$.

There are many different parametric approaches to modelling the news impact curve and they can give quite different answers in the range of perhaps most interest to practitioners. This motivates a nonparametric approach, because of the greater flexibility in functional form thereby allowed. The nonparametric ARCH literature apparently begins with Pagan and Schwert (1990) and Pagan and Hong (1991). They consider the case where $\sigma_t^2 = \sigma^2(y_{t-1})$, where $\sigma(\cdot)$ is a smooth but unknown function, and the multilag version $\sigma_t^2 = \sigma^2(y_{t-1}, y_{t-2}, \ldots, y_{t-d})$. This allows for a general shape to the news impact curve and nests all the parametric ARCH processes. Under some general conditions on $\sigma(\cdot)$ (for example that $\sigma(\cdot)$ does not grow at a more than quadratic rate in the tails) the process y is geometrically strong mixing. Härdle and Tsybakov (1997) applied local linear fit to estimate the volatility function together with the mean function and derived their joint asymptotic properties. The multivariate extension is given in Härdle, Tsybakov and Yang (1996). Masry and Tjøstheim (1995) also estimate nonparametric ARCH models using the Nadaraya-Watson kernel estimator. Lu and Linton (2006) extend the CLT to processes that are only near epoch dependent. Fan and Yao (1998) have discussed efficiency issues in this model, see also Avramidis (2002). Franke, Neumann, and Stockis (2004) have considered the application of bootstrap for improved inference. In practice, it is necessary to include many lagged variables in $\sigma^2(.)$ to match the dependence found in financial data. The problem with this is that nonparametric estimation of a multi-dimension regression surface suffers from the well-known "curse of dimensionality": the optimal rate of convergence decreases with dimensionality d, see Stone (1980). In addition, it is hard to describe, interpret and understand the estimated regression surface when the dimension is more than two. Furthermore, even for large d this model greatly restricts the dynamics for the variance process since it effectively corresponds to an ARCH(d) model, which is known in the parametric case not to capture the dynamics well. In particular, if the conditional variance is highly persistent, the non-parametric estimator of the conditional variance will provide a poor approximation, as reported in Perron (1998). So not only does this model not capture adequately the time series properties of many datasets, but the statistical properties of the estimators can be poor, and the resulting estimators hard to interpret.

Additive models offer a flexible but parsimonious alternative to nonparametric models, and have been used in many contexts, see Hastie and Tibshi-

rani (1990). Suppose that

$$\sigma_t^2 = c_v + \sum_{j=1}^{d} \sigma_j^2(y_{t-j}) \tag{5}$$

for some unknown functions σ_j^2. The functions σ_j^2 are allowed to be of general functional form but only depend on y_{t-j}. This class of processes nests many parametric ARCH models. Again, under growth conditions the process y can be shown to be stationary and geometrically mixing. The functions σ_j^2 can be estimated by special kernel regression techniques, such as the method of marginal integration, see Linton and Nielsen (1995) and Tjøstheim and Auestad (1994). The best achievable rate of convergence for estimates of $\sigma_j^2(.)$ is that of one-dimensional nonparametric regression, see Stone (1985). Masry and Tjøstheim (1995) develop estimators for a class of time series models including (5). Yang, Härdle, and Nielsen (1999) proposed an alternative nonlinear ARCH model in which the conditional mean is again additive, but the volatility is multiplicative $\sigma_t^2 = c_v \prod_{j=1}^{d} \sigma_j^2(y_{t-j})$. Kim and Linton (2004) generalize this model to allow for arbitrary [but known] transformations, i.e., $G(\sigma_t^2) = c_v + \sum_{j=1}^{d} \sigma_j^2(y_{t-j})$, where $G(.)$ is a known function like log or level. The typical empirical findings are that the news impact curves have an inverted asymmetric U-shape.

These models address the curse of dimensionality but they are rather restrictive with respect to the amount of information allowed to affect volatility, and in particular do not nest the GARCH(1,1) process. Linton and Mammen (2005) proposed the following model

$$\sigma_t^2(\theta, m) = \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j}), \tag{6}$$

where $\theta \in \Theta \subset \mathbb{R}^p$ and m is an unknown but smooth function. The coefficients $\psi_j(\theta)$ satisfy at least $\psi_j(\theta) \geq 0$ and $\sum_{j=1}^{\infty} \psi_j(\theta) < \infty$ for all $\theta \in \Theta$. A special case of this model is the Engle and Ng (1993) PNP model where

$$\sigma_t^2 = \beta \sigma_{t-1}^2 + m(y_{t-j}),$$

where $m(.)$ is a smooth but unknown function. This model nests the simple GARCH (1,1) model but permits more general functional form: it allows for an asymmetric leverage effect, and as much dynamics as GARCH(1,1). Estimation methods for these models are based on iterative smoothing. Linton and Mammen (2005) show that the news impact curves for daily and weekly S&P500 data are quite asymmetric with non-quadratic tails and is not minimal at zero but at some positive return. Below we show their estimator, denoted PNP here, in comparison with a common parametric fit, denoted AGARCH.

Yang (2006) introduced a semiparametric index model

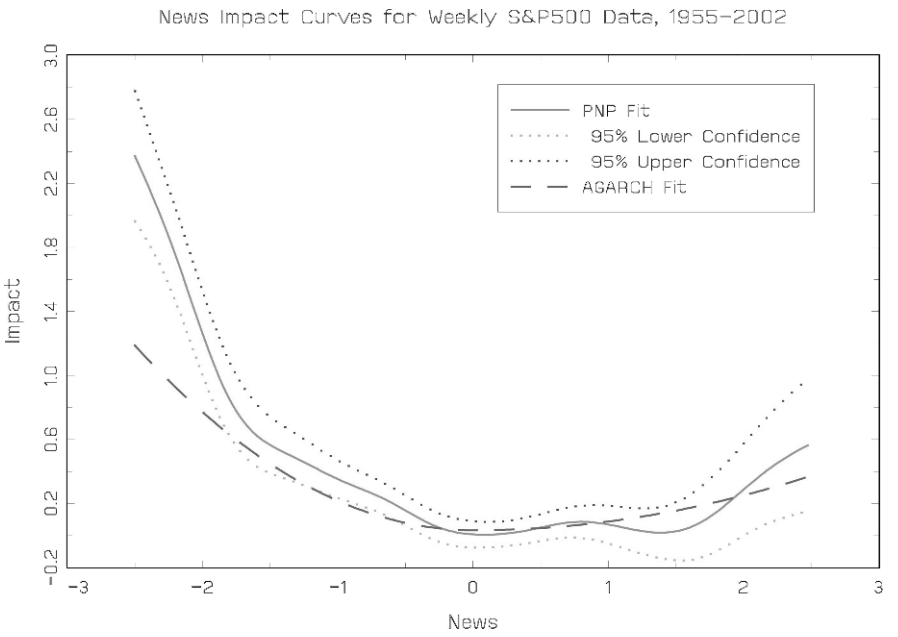

Fig. 1 Partially nonparametric (PNP) (with pointwise confidence band) and parametric fit of news impact curve for weekly Standard & Poors 500 data.

$$\sigma_t^2 = g\left(\sum_{j=1}^{\infty} \nu_j(y_{t-j};\theta)\right),$$

where $\nu_j(y;\theta)$ are known functions for each j satisfying some decay condition and g is smooth but unknown. This process nests the GARCH(1,1) when g is the identity, but also the quadratic model considered in Robinson (1991).

Finally, we should mention some work by Audrino and Bühlmann (2001): their model is that $\sigma_t^2 = \Lambda(y_{t-1}, \sigma_{t-1}^2)$ for some smooth but unknown function $\Lambda(.)$, and includes the PNP model as a special case. They proposed an estimation algorithm. However, they did not establish the distribution theory of their estimator, and this may be very difficult to establish due to the generality of the model.

3.3 Relationship between mean and variance

The above discussion has centered on the evolution of volatility itself, whereas one is often very interested in the mean as well. One might expect that risk and return should be related, Merton (1973). The GARCH-in-Mean process

captures this idea, it is

$$y_t = g(\sigma_t^2; b) + \varepsilon_t \sigma_t,$$

for various functional forms of g e.g., linear and log-linear and for some given specification of σ_t^2. Engle, Lilien and Robbins (1987) introduced this model and applied it to the study of the term Structure. Here, b are parameters to be estimated along with the parameters of the error variance. Some authors find small but significant effects. Again, the nonparametric approach is well motivated here on grounds of flexibility. Pagan and Hong (1991) and Pagan and Ullah (1988) consider a case where the conditional variance is nonparametric (with a finite number of lags) but enters in the mean equation linearly or log linearly. Linton and Perron (2002) studied the case where g is nonparametric but σ_t^2 is parametric, for example GARCH. The estimation algorithm was applied to stock index return data. Their estimated g function was non-monotonic for daily S&P500 returns.

3.4 Long memory

Another line of work has argued that conventional models involve a dependence structure that does not fit the data well enough. The GARCH(1,1) process $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2$ is of the form

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j y_{t-j}^2 \tag{7}$$

for constants c_j satisfying $c_j = \gamma\beta^{j-1}$, provided the process is weakly stationary, which requires $\gamma + \beta < 1$. These coefficients decay very rapidly so the actual amount of memory is quite limited. There is some empirical evidence on the autocorrelation function of y_t^2 for high frequency returns data that suggests a slower decay rate than would be implied by these coefficients, see Bollerslev and Mikkelson (1996). Long memory models essentially are of the form (7) but with slower decay rates. For example, suppose that $c_j = j^{-\theta}$ for some $\theta > 0$. The coefficients satisfy $\sum_{j=1}^{\infty} c_j^2 < \infty$ provided $\theta > 1/2$. Fractional integration (FIGARCH) leads to such an expansion. There is a single parameter called d that determines the memory properties of the series, and

$$(1-L)^d \sigma_t^2 = \omega + \gamma\sigma_{t-1}^2(\varepsilon_{t-1}^2 - 1),$$

where $(1-L)^d$ denotes the fractional differencing operator. When $d = 1$ we have the standard IGARCH model. For $d \neq 1$ we can define the binomial expansion of $(1-L)^{-d}$ in the form given above. See Robinson (1991) and Bollerslev and Mikkelson (1996) for models. The evidence for long memory is often based on sample autocovariances of y_t^2, and this may be questionable

when only few moments of y_t exist, see Mikosch and Stărică (2002). See Giraitis et al. (2008) for a nice review.

3.5 Locally stationary processes

Recently, another criticism of GARCH processes has come to the fore, namely their usual assumption of stationarity. The IGARCH process (where $\beta+\gamma = 1$) is one type of nonstationary GARCH model but it has certain undesirable features like the non-existence of the variance. An alternative approach is to model the coefficients of a GARCH process as changing over time, thus

$$\sigma_t^2 = \omega(x_{tT}) + \beta(x_{tT})\sigma_{t-1}^2 + \gamma(x_{tT})(y_{t-1} - \mu_{t-1})^2,$$

where ω, β, and γ are smooth but otherwise unknown functions of a variable x_{tT}. When $x_{tT} = t/T$, this class of processes is nonstationary but can be viewed as locally stationary along the lines of Dahlhaus (1997), provided the memory is weak, i.e., $\beta(\cdot) + \gamma(\cdot) < 1$. In this way the unconditional variance exists, i.e., $E[\sigma_t^2] < \infty$, but can change slowly over time as can the memory. Engle and Rangel (2006) impose some restrictions that makes the unconditional variance $\sigma^2(t/T) = \omega(t/T)/(1 - \beta(t/T) - \gamma(t/T))$ vary over time but the coefficients $\beta(t/T)$ and $\gamma(t/T)$ are assumed to be constant. Dahlhaus and Subba Rao (2006) have recently provided a comprehensive theory of such processes and about inference methods for the ARCH special case. See Čížek and Spokoiny (2008) for a further review.

3.6 Continuous time

Recently there has been much work on nonparametric estimation of continuous time processes, see for example Bosq (1998). Given a complete record of transaction or quote prices, it is natural to model prices in continuous time (e.g., Engle (2000)). This matches with the vast continuous time financial economic arbitrage-free theory based on a frictionless market. Under the standard assumptions that the return process does not allow for arbitrage and has a finite instantaneous mean, the asset price process, as well as smooth transformations thereof, belong to the class of special semi-martingales, as detailed by Back (1991). Under some conditions, the semiparametric GARCH processes we reviewed can approximate such continuous time processes as the sampling interval increases. Work on continuous time is reviewed elsewhere in this volume, so here we just point out that this methodology can be viewed as nonparametric and as a competitor of the discrete time models we outlined above.

4 Conclusion

In conclusion, there have been many advances in the application of nonparametric methods to the study of volatility, and many difficult problems have been overcome. These methods have offered new insights into functional form, dependence, tail thickness, and nonstationarity that are fundamental to the behaviour of asset returns. They can be used by themselves to estimate quantities of interest like value at risk. They can also be used as a specification device enabling the practitioner to see with respect to which features of the data their parametric model is a good fit.

References

Audrino, F. and Bühlmann, P (2001): Tree-structured GARCH models. *Journal of The Royal Statistical Society* **63**, 727–744.

Avramidis, P. (2002): Local maximum likelihood estimation of volatility function. *Manuscript, LSE.*

Back, K. (1991): Asset Pricing for General Processes. *Journal of Mathematical Economics* **20**, 371–395.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T., Engle, R. F. and Nelson, D. B (1994): ARCH Models. *In: Engle, R.F. and McFadden, D.L. (Eds.): Handbook of Econometrics* **IV**, 2959-3038. Elsevier Science.

Bosq, D. (1998): *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction.* Springer, Berlin.

Carrasco, M. and Chen, X. (2002): Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models. *Econometric Theory* **18**, 17–39.

Čížek, P. and Spokoiny, V. (2008): Varying coefficient GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 168–185. Springer, New York.

Dahlhaus, R. and Subba Rao, S. (2006): Statistical inference for time-varying ARCH processes. *The Annals of Statistics* **34**, 1075–1114.

Drost, F.C. and Klaassen, C.A.J. (1997): Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* **81**, 193–221.

Engle, R.F. (1982): Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987–1008.

Engle, R.F. (2000): The Econometrics of Ultra-High Frequency Data. *Econometrica* **68**, 1–22.

Engle, R.F. and González-Rivera, G. (1991): Semiparametric ARCH models. ßsl Journal of Business and Economic Statistics **9**, 345–359.

Engle, R.F., Lilien, D.M. and Robins, R.P. (1987): Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model. *Econometrica* **19**, 3–29.

Engle, R.F. and Ng, V.K. (1993): Measuring and Testing the impact of news on volatility. *The Journal of Finance* **XLVIII**, 1749–1778.

Engle, R.F. and Rangel, J.G. (2006): The spline GARCH model for unconditional volatility and its global macroeconomic causes. *Mimeo, NYU.*

Fama, E.F. (1965): The behavior of stock market prices. *Journal of Business* **38**, 34–105.

Fan, J. and Yao, Q. (1998): Efficient estimation of conditional variance functions in Stochastic Regression. *Biometrika* **85**, 645–660.

Franke, J., Neumann, M.H. and Stockis, J. (2004): Bootstrapping nonparametric estimators of the volatility function. *Journal of Econometrics* **118**, 189–218.

Giraitis, L., Leipus, R. and Surgailis, D. (2008): ARCH(∞) and long memory properties. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 70–84. Springer, New York.

Glosten, L. R., Jagannathan, R. and Runkle, D. E: (1993): On the Relation Between the Expected Value and the Volatility of the Nominal Excess Returns on Stocks. *Journal of Finance* **48**, 1779–1801.

Hafner, C.M. and Rombouts, J.V.K (2006): Semiparametric Multivariate Volatility Models. *Econometric Theory*, to appear.

Härdle, W. and Tsybakov, A.B. (1997): Local polynomial estimators of the volatility function. *Journal of Econometrics* **81**, 223–242.

Kim, W. and Linton, O.B. (2004): A Local Instrumental Variable Estimation method for Generalized Additive Volatility Models. *Econometric Theory* **20**, 1094–1139..

Linton, O.B. (1993): Adaptive estimation in ARCH models. *Econometric Theory* **9**, 539–569.

Linton, O.B. and Mammen, E. (2005): Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* **73**, 771–836.

Linton, O.B. and Nielsen, J.P. (1995): A kernel method of estimating structured nonparametric regression using marginal integration. *Biometrika* **82**, 93–100.

Linton, O. B. and Perron, B. (2003): The Shape of the Risk Premium: Evidence from a Semiparametric Generalized Autoregressive Conditional Heteroskedasticity Model. *Journal of Business & Economic Statistics*, 354–367.

Lu, Z. and Linton, O.B. (2006): Local linear fitting under Near Epoch Dependence. *Econometric Theory*, to appear.

Mammen, E., Linton, O.B. and Nielsen, J.P. (1999): The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27**, 1443–1490.

Mandelbrot, B. (1963): The variation of certain speculative prices. *Journal of Business* **36**, 394–419.

Masry, E. and Tjøstheim, D. (1995): Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory* **11**, 258–289.

Merton, R.C. (1973): An Intertemporal Capital Asset Pricing Model. *Econometrica* **41**, 867–887.

Mikosch, T. and Stărică, C. (2000): Limit Theory for the Sample Autocorrelations and Extremes of a GARCH(1,1) Process. *Annals of Statistics* **28**, 1427–1451.

Nelson, D.B. (1990): Stationarity and Persistence in the GARCH(1,1) model. *Econometric Theory* **6**, 318–34.

Nelson, D.B. (1991): Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59**, 347–370.

Newey, W. K. and Steigerwald, D.G. (1997): Asymptotic Bias for Quasi-Maximum-Likelihood Estimators in Conditional Heteroskedasticity Models. *Econometrica* **65**, 587–599.

Pagan, A.R. and Schwert, G.W. (1990): Alternative models for conditional stock volatility. *Journal of Econometrics* **45**, 267–290.

Pagan, A.R. and Hong, Y.S. (1991): Nonparametric Estimation and the Risk Premium. *In: Barnett, W., Powell, J. and Tauchen, G.E. (Eds.): Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press.

Robinson, P.M. (1991): Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* **47**, 67–84.

Robinson, P.M. and Zaffaroni, P. (2006): Pseudo-maximum likelihood estimation of ARCH(∞) models. *Annals of Statistics* **34**, 1049–1074.

Stone, C.J. (1980): Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* **8**, 1348–1360.

Stone, C.J. (1985): Additive regression and other nonparametric models. *Annals of Statistics* **13**, 685–705.

Teräsvirta, T., Tjøstheim, D. and Granger, C.W.J. (1994): Aspects of Modelling Nonlinear Time Series. *In: McFadden, D.L. and Engle, R.F. (Eds.): The Handbook of Econometrics* **IV**, 2919–2960. Elsevier, Amsterdam.

Tjøstheim, D. and Auestad, B. (1994): Nonparametric identification of nonlinear time series: projections. *J. Am. Stat. Assoc.* **89**, 1398–1409.

Tong, H. (1990): *Nonlinear Time Series Analysis: A dynamical systems Approach*. Oxford University Press, Oxford.

Wu, G. and Xiao, Z. (2002): A generalized partially linear model for asymmetric volatility. *Journal of Empirical Finance* **9**, 287–319.

Yang, L. (2006): A Semiparametric GARCH Model for foreign exchange volatility. *Journal of Econometrics* **130**, 365–384.

Yang, L., Härdle, W. and Nielsen, J.P. (1999): Nonparametric Autoregression with Multiplicative Volatility and Additive Mean. *Journal of Time Series Analysis* **20**, 579–604.

Ziegelmann, F. (2002): Nonparametric estimation of volatility functions: the local exponential estimator. *Econometric Theory* **18**, 985–992.

Varying Coefficient GARCH Models

Pavel Čížek and Vladimir Spokoiny

Abstract This paper offers a new method for estimation and forecasting of the volatility of financial time series when the stationarity assumption is violated. We consider varying–coefficient parametric models, such as ARCH and GARCH, whose coefficients may arbitrarily vary with time. This includes global parametric, smooth transition, and change–point models as special cases. The method is based on an adaptive pointwise selection of the largest interval of homogeneity with a given right–end point, which is obtained by a local change–point analysis. We construct locally adaptive volatility estimates that can perform this task and investigate them both from the theoretical point of view and by Monte Carlo simulations. Additionally, the proposed method is applied to stock–index series and shown to outperform the standard parametric GARCH model.

1 Introduction

A growing amount of econometrical and statistical research is devoted to modeling financial time series and their volatility, which measures dispersion at a point in time (e.g., conditional variance) and which is one of crucial quantities in risk management and derivative pricing. Although financial markets have been recently experiencing many shorter and longer periods of instability or uncertainity such as Asian crisis in 1997, Russian crisis in 1998, start of the European currency in 1999, the "dot–Com" technology–bubble

Pavel Čížek
Department of Econometrics & OR, Tilburg University, Tilburg, P.O.Box 90153, 5000 LE Tilburg, The Netherlands, e-mail: P.Cizek@uvt.nl

Vladimir Spokoiny
Weierstrass–Institute, Mohrenstrasse 39, D–10117 Berlin, Germany, e-mail: spokoiny@wias-berlin.de

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_7,

crash (2000–2002), or the terrorist attacks (September, 2001) and the war in Iraq (2003), mostly used econometric models are typically based on the assumption of time homogeneity. This includes conditional heteroscedasticity models such as ARCH (cf. Engle (1982)) and GARCH (cf. (Bollerslev (1986)), stochastic volatility models (Taylor (1986)), and many of their descendants. On the other hand, the market and institutional changes have long been assumed to cause structural breaks in financial time series, which was confirmed in stock–price and exchange–rate series, for example, by Andreou and Ghysels (2002) and Herwatz and Reimers (2001), respectively. Moreover, ignoring these breaks can adversely affect the modeling, estimation, and forecasting of volatility as suggested by Diebold and Inoue (2001), Mikosch and Stărică (2004), Pesaran and Timmermann (2004), and Hillebrand (2005), for instance. Such findings led to the development of the change–point analysis in the context of conditional heteroscedasticity models; see for example, Chu (1995), Chen and Gupta (1997), Lin and Yang (2000), Kokoszka and Leipus (2000), or Andreou and Ghysels (2006).

An alternative approach lies in relaxing the assumption of time homogeneity and allowing some or all model parameters to vary over time (Fan and Zhang (1999); Cai et al. (2000); Fan et al. (2003)). Without structural assumptions about the transition of model parameters over time, time–varying models have to be estimated nonparametrically, for example, under the identification condition that their parameters are smooth functions of time. In this chapter, we follow a more general strategy based on the assumption that a time series can be locally, that is over short periods of time, approximated by a parametric model. As suggested by Spokoiny (1998), such a local approximation can form a starting point in the search for the longest period of stability (homogeneity), that is, for the longest time interval in which the series is described by the given parametric model. This strategy in the context of the local constant approximation of the volatility process was employed for by Härdle et al. (2003), Mercurio and Spokoiny (2004), and Spokoiny and Chen (2007).

Here we generalize the method of Mercurio and Spokoiny (2004) so that it can identify intervals of homogeneity for more complex parametric model of volatility. The main benefit of such a generalization consists in the possibility to forecast over a longer time horizon: the assumption of a constant volatility is fulfilled only for short time intervals, whereas parametric models like ARCH and GARCH mimic the majority of stylized facts about financial time series and can reasonably fit the data over rather long periods of time in many practical situations. Allowing for time dependence of model parameters offers then much more flexibility in modeling the real–life financial time series, which can be both with or without structural breaks since global parametric models are included as a special case.

Moreover, the proposed adaptive local parametric modeling unifies the change–point and varying–coefficient approaches. First, since finding the longest time–homogeneous interval for a parametric model at any point in

time corresponds to detecting the most recent change–point in a time series, this approach is analogous to the change–point modeling as in Mikosch and Stărică (1999, 2004), for instance, but it does not require prior information such as the number of changes or the minimal time interval between two neighboring changes. Second, since the selected interval used for estimation necessarily differs at each time point, the model coefficients can vary arbitrarily over time. In comparison to varying–coefficient models assuming smooth development of parameters over time (Cai et al. (2000)), our approach however allows for structural breaks in the form of sudden jumps in parameter values.

The rest of the chapter is organized as follows. In Section 2, we discuss the two main ingredients of the method: parameter estimation of conditional heteroscedasticity models and the test of homogeneity for these models. Section 3 introduces the adaptive estimation procedure and discusses the choice of its parameters. The performance of the proposed methods is demonstrated on an application to real stock–index series in Section 4.

2 Conditional Heteroscedasticity Models

Let S_t be an observed asset–price process in discrete time, $t = 1, 2, \ldots$, and R_t are the corresponding returns, $R_t = \log(S_t/S_{t-1})$. We model this process via the conditional heteroscedasticity assumption

$$R_t = \sigma_t \varepsilon_t \,, \tag{1}$$

where $\{\varepsilon_t\}_{t\geq 1}$ is a sequence of independent standard Gaussian random variables and σ_t is the volatility process.

Standard ways of modeling the volatility process σ_t rely on one or another parametric assumption. A large class of parametric models is built around the ARCH (Engle (1982)) and GARCH (Bollerslev (1986)) models. The ARCH(p) model is described by the equation

$$\sigma_t^2 = \omega + \alpha_1 R_{t-1}^2 + \ldots + \alpha_p R_{t-p}^2.$$

Its main disadvantage is that the order p of this autoregressive type equation has to be relatively large in order to capture the main stylized facts of the financial time series. The more general GARCH(p, q) model is described by an ARMA–type equation

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i R_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \tag{2}$$

An attractive feature of this model is that, even with very few coefficients, one can model most stylized facts of financial time series like volatility clustering or excessive curtosis, for instance. A number of GARCH extensions were proposed to make the model even more flexible; for example, EGARCH (Nelson (1991)), QGARCH (Sentana (1995)), TGARCH (Glosten, Jagannathan, and Runkle (1993)), and APARCH (Ding, Granger, and Engle (1993)) that account for asymmetries in a volatility process and (fractionally) integrated versions of GARCH (Nelson (1990); Baillie, Bollerslev, and Mikkelsen (1996)). Further developments include stochastic volatility models which include an additional random noise component on the right–hand side of (2) or its alternatives.

All these models can be put into a common class of generalized linear volatility models:

$$R_t \sim \mathcal{N}(0, \sigma_t^2), \quad \sigma_t^2 = g(X_t), \quad X_t = \omega + \sum_{i=1}^{p} \alpha_i h(R_{t-i}) + \sum_{j=1}^{q} \beta_j X_{t-j}, \tag{3}$$

where g and h are known functions and X_t is a latent process (structural variable) that models volatility coefficient σ_t^2 via transformation g. For example, we mostly concentrate on the GARCH case, which is described by $g(u) = u$ and $h(r) = r^2$. In what follows, we denote $Y_t = h(R_t)$ and write the model equation in the linear form

$$X_t = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j X_{t-j}. \tag{4}$$

In particular, the GARCH(1,1) model reads as $X_t = \omega + \alpha Y_{t-1} + \beta X_{t-1}$. Usually all the coefficients are assumed nonnegative, that is, $\omega \geq 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. The condition $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$ ensures ergodicity of the process X_t.

Model (3) is time–homogeneous in the sense that the volatility process follows the same structural equation at each time point. In other words, parameters $\omega, \{\alpha_i\}_{i=1}^{p}$, and $\{\beta_j\}_{j=1}^{q}$ are constant over time. Even though the conditional heteroscedasticity models can often fit data well over a longer period of time, the assumption of homogeneity is too restrictive in practical applications: to guarantee sufficient amount of data for sufficiently precise estimation, these models are often applied for time spans of many years. The strategy pursued in this paper requires only local time homogeneity, which means that at each time point t there is a (possibly rather short) interval $[t - m, t]$, where the volatility process σ_t is well described by model (3). This strategy aims then both at finding an interval of homogeneity (preferably as long as possible) and at the estimation of the corresponding value σ_t.

To facilitate such a time–adaptive estimation, the estimation of model (3) (Section 2.1) and a test of homogeneity of a given time interval (Section 2.2) have to be described in more details.

2.1 Model estimation

This section discusses the parameter estimation for the conditional heteroscedasticity model (3) using the observations R_t from some time interval $I = [t_0, t_1]$.

The volatility process σ_t^2 is described by equation $\sigma_t^2 = g(X_t)$, where the structural process X_t fulfills the linear constraint (3). Neglecting the (typically small) influence of initial conditions in the ergodic case, the process at time t is determined by the parameter vector $\boldsymbol{\theta} = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^\top$ and past values of X_t and Y_t. Therefore, we use notation $X_t = X_t(\boldsymbol{\theta})$,

$$X_t(\boldsymbol{\theta}) = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j X_{t-j}(\boldsymbol{\theta}), \tag{5}$$

for $t \in I$. Later, we additionally use symbol $\boldsymbol{\theta}(t)$ to indicate the dependence of parameters on time t.

For estimating the parameter $\boldsymbol{\theta}$, we apply the quasi maximum likelihood (quasi–MLE) approach which guarantees efficiency under the normality of innovations and consistency under rather general moment conditions (e.g., Hansen and Lee (1994)). The quasi log–likelihood for model (3) on an interval I can be represented in the form

$$L_I(\boldsymbol{\theta}) = \sum_{t \in I} \ell\big(R_t, g\{X_t(\boldsymbol{\theta})\}\big)$$

with $\ell(r, u) = -0.5\big\{\log(u) + r^2/u\big\}$. We define the quasi–MLE estimate $\tilde{\boldsymbol{\theta}}_I$ of the parameter $\boldsymbol{\theta}$ by maximizing the log–likelihood $L_I(\boldsymbol{\theta})$,

$$\tilde{\boldsymbol{\theta}}_I = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L_I(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{t \in I} \ell\big(R_t, g\{X_t(\boldsymbol{\theta})\}\big), \tag{6}$$

and denote by $L_I(\tilde{\boldsymbol{\theta}}_I)$ the corresponding maximum.

2.2 Test of homogeneity against a change–point alternative

Using interval $I = [t_0, t_1]$ further, we want to test now whether the observed returns R_t follow a parametric model (3) within I. We consider the supremum likelihood ratio (LR) test introduced by Andrews (1993) against a change–point alternative within I. Although in the context of conditional heteroscedasticity models, there are other tests of homogeneity against a change–point alternative (e.g., Kokozska and Leipus (1999, 2000)), the LR–type tests are preferable both from the theoretical perspective (development

of theory) and practical point of view (performance in finite samples; see Kokoszka and Teyssiere (2005).

The null hypothesis for I means that the observations $\{R_t\}_{t\in I}$ follow the parametric model (3) with a parameter $\boldsymbol{\theta}(t) = \boldsymbol{\theta}^*$. The null hypothesis yields the parametric estimate $\tilde{\boldsymbol{\theta}}_I$ due to (6) and the corresponding fitted log–likelihood $L_I(\tilde{\boldsymbol{\theta}}_I)$. The change–point alternative given by a set $\mathcal{T}(I)$ of possible change points within I can be described as follows. Every point $\tau \in \mathcal{T}(I)$ splits the interval I in two subintervals $J = [t_0, \tau]$ and $J^c = I \setminus J = [\tau+1, t_1]$. A change–point alternative with a location at $\tau \in \mathcal{T}(I)$ means that $\boldsymbol{\theta}(t) = \boldsymbol{\theta}_1$ for $t \in J$ and $\boldsymbol{\theta}(t) = \boldsymbol{\theta}_2$ for $t \in J^c$ with two different values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Under such an alternative, the observations $\{R_t\}_{t\in I}$ are associated with the log–likelihood $L_J(\boldsymbol{\theta}_1) + L_{J^c}(\boldsymbol{\theta}_2)$.

To test against a single change–point alternative with a known fixed $\tau \in \mathcal{T}(I)$, the LR test statistic can be be used:

$$\begin{aligned} T_{I,\tau} &= 2[\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Theta} \{L_J(\boldsymbol{\theta}_1) + L_{J^c}(\boldsymbol{\theta}_2)\} - \max_{\boldsymbol{\theta}\in\Theta} L_I(\boldsymbol{\theta})] \\ &= 2[L_J(\tilde{\boldsymbol{\theta}}_J) + L_{J^c}(\tilde{\boldsymbol{\theta}}_{J^c}) - L_I(\tilde{\boldsymbol{\theta}}_I)]. \end{aligned}$$

Considering an unknown change point $\tau \in \mathcal{T}(I)$, the change–point test for interval I can be defined as the maximum (supremum) of the LR statistics over all $\tau \in \mathcal{T}(I)$:

$$T_I = \max_{\tau\in\mathcal{T}(I)} T_{I,\tau}\,. \tag{7}$$

A change point is detected within I if the test statistic T_I exceeds a critical value $\mathfrak{z}$.

The supremum LR test was proposed and studied by Andrews (1993). To guarantee the convergence to a parameter–independent asymptotic distribution, the distance between $\mathcal{T}(I)$ and the end points of interval I has to increase with the interval length $|I| = t_1 - t_0 + 1$. Specifically, it should hold for some $\delta > 0$ that

$$\min\{t_1 - \max\{\tau : \tau \in \mathcal{T}(I)\}, \min\{\tau : \tau \in \mathcal{T}(I)\} - t_0\} > \delta|I|. \tag{8}$$

This asymptotic result is however almost useless for our purposes because the proposed adaptive procedure applies such a change–point test for very short intervals (even less than 20 observations). The choice of finite–sample critical values for this test statistics will be further discussed in Sections 3 and critical values specific to the ARCH(1) and GARCH(1,1) models will be obtained in Section 4.

3 Adaptive Nonparametric Estimation

An obvious feature of model (3) is that the parametric structure of the process is assumed constant over the whole sample and cannot thus incorporate changes and structural breaks in the model. A natural generalization leads to models whose coefficients may vary with time. Cai et al. (2000) considered the following varying–coefficient model

$$R_t \sim \mathcal{N}(0, \sigma_t^2), \quad \sigma_t^2 = g(X_t), \quad X_t = \omega(t) + \sum_{i=1}^{p} \alpha_i(t) Y_{t-i} + \sum_{j=1}^{q} \beta_j(t) X_{t-j},$$

where $\omega(t)$, $\alpha_i(t)$, and $\beta_j(t)$ are functions of time and have to be estimated from the observations R_t. Naturally, this is only possible under some additional assumptions on these functions, which are typically (i) varying coefficients are smooth functions of time (Cai et al. (2000)) and (ii) varying coefficients are piecewise constant (piecewise smooth) functions (Mikosch and Stărică (1999, 2004)).

Our approach is based on a more general structural assumption: at each point T, there exists an interval of homogeneity $I(T) = [t, T]$ in which the volatility X_t nearly follows the parametric conditional heteroscedasticity specification (4). This particularly means that, within the interval $I(T)$, the returns $\{R_t\}_{t \in I(T)}$ can be assumed to be driven by the parametric model (3) with some parameter $\boldsymbol{\theta} = \boldsymbol{\theta}(T)$. The adaptive pointwise method applied here permits selecting $I(T)$ and the corresponding $\boldsymbol{\theta}(T)$ independently at every point T, which allows us to incorporate the mentioned special cases (i) and (ii) in a unified way. Specifically, the idea of the method is to find the longest interval, where the data do not contradict the hypothesis of model homogeneity. Starting at each time T with a very short interval, the search is done by successive extending and testing of interval $I(T)$ on homogeneity against a change–point alternative. If the hypothesis of homogeneity is not rejected for a given $I(T)$, a larger interval is taken and tested again. Hence, the main difference between this method and that of Mikosch and Stărică (1999) is that we do not aim to detect all change points in a given time series but focus on the local change–point analysis near the point of estimation. This allows restricting to local alternatives with only one change point.

In the rest of this section, the adaptive pointwise estimation procedure is rigorously described (Section 3.1). Further, the choice of parameters and implementation of the adaptive procedure is described in Section 3.2 (in particular, the choice of critical values is dealt with in Section 3.2.2).

3.1 Adaptive choice of the interval of homogeneity

This section proposes an adaptive method for the pointwise selection of the longest interval of homogeneity for the conditional heteroscedastic model (3). Let T be the point of interest and let $\mathcal{I}$ be a sequence of intervals $[t_k, T], k = 0, \ldots, K$, defined by time points $T > t_0 > \ldots > t_K = 1$ (specific proposals for the choice of these points are discussed later in this section). We aim to successively test every interval $I_k = [t_k, T] \in \mathcal{I}$ on homogeneity, where k increases from 2 to K. This ensures that we find the most recent change point, and consequently, that we can search only one change point at a time. Additionally, sets $\mathcal{T}(I_k)$ of possible change–points that we test for within I_k are defined such that $\mathcal{T}(I_k)$ does not intersect with I_{k-2}, $\mathcal{T}(I_k) \cap I_{k-2} = \emptyset$, which can be achieved by setting $\mathcal{T}(I_k) = I_{k-1} \backslash I_{k-2}$, for instance. Therefore, at the step k, we check for a possible change only points which have not yet been tested at previous steps. The rejection of the time–homogeneity hypothesis in I_k then means a detection of a change point and rejection of the time-homogeneity of I_k, and at the same time, acceptance of I_{k-2} as the longest time–homogeneous interval.

The procedure reads as follows.

Initialization. Set $k = 2$.

Iteration. Select interval $I_k = [t_k, T]$.

Testing homogeneity. Test the hypothesis of homogeneity within I_k against a change-point alternative as described in Section 2.2 by comparing the test statistic T_{I_k} with the critical value $\mathfrak{z}_k$.

Loop. If a change point is detected for I_k, then set $\hat{I}_k = I_{k-2} = [t_{k-2}, T]$, set $\hat{I}_l = \hat{I}_k, l = k+1, \ldots, K$, and terminate. Otherwise, accept $\hat{I}_k = I_k$, increase k by one, and continue with the iteration step until $k > K$.

In the description of the adaptive procedure above, $\hat{I}_k$ denotes the latest accepted interval after the first k steps of the procedure. The corresponding quasi–MLE estimate on $\hat{I}_k$ is then $\hat{\boldsymbol{\theta}}_k = \tilde{\boldsymbol{\theta}}_{\hat{I}_k}$. The final adaptively selected interval $\hat{I}(T)$ at T is the latest accepted interval from the whole procedure, that is, $\hat{I}(T) = \hat{I}_K$. The corresponding adaptive pointwise estimate $\hat{\boldsymbol{\theta}}(T)$ of $\boldsymbol{\theta}(T)$ is then defined as $\hat{\boldsymbol{\theta}}(T) = \tilde{\boldsymbol{\theta}}_{\hat{I}(T)}$.

3.2 Parameters of the method and the implementation details

To run the proposed procedure, one has to fix some of its parameters and ingredients. This especially concerns the choice of tested intervals I_k, and for every I_k, a subset $\mathcal{T}(I_k)$ of tested change point locations within I_k (Section 3.2.1). The most important ingredient of the method is a collection of the

critical values $\mathfrak{z}_k$. Their choice and related computational issues are discussed in Sections 3.2.2–3.2.3.

3.2.1 Set of intervals

This section presents our way of selecting the sets I_k and $\mathcal{T}(I_k)$ for $k = 1, \ldots, K$. Note however that our proposal is just an example and the method will apply under rather general conditions on these sets. In what follows, similarly to Spokoiny and Chen (2007), we fix some geometric grid $\{m_k = [m_0 a^k], k = 0, \ldots, K\}$ with an initial length $m_0 \in N$ and a multiplier $a > 1$ to define intervals $I_k = [t_k, T] = [T - m_k, T]$, $k = 0, \ldots, K$. For every interval I_k, the subset $\mathcal{T}(I_k)$ of considered change point locations is then defined as $I_{k-1} \backslash I_{k-2}$, which guarantees that $\mathcal{T}(I_k) \cap \mathcal{T}(I_l) = \emptyset$ for $l \neq k$. Note that all intervals depend on the reference end point T.

Our experiments show that results are rather insensitive to the choice of the parameters a and m_0. Results presented in Section 4 employ a multiplier $a = 1.25$ and the initial length $m_0 = 10$.

3.2.2 Choice of the critical values $\mathfrak{z}_k$

The proposed estimation method can be viewed as a hierarchic multiple testing procedure. The parameters $\mathfrak{z}_k$ are thus selected to provide the prescribed error level under the null hypothesis, that is, in the parametric time–homogeneous situation. Because the proposed adaptive choice of the interval of homogeneity is based on the supremum LR test applied sequentially in rather small samples, the asymptotic properties of the supremum LR statistics T_I defined in (7) for a single interval I (Andrews (1993)) are not applicable. Instead of asymptotic bounds, we therefore choose the critical values using the Monte–Carlo simulations and the theoretical concept presented in this section.

In what follows we assume that the considered set $\mathcal{I}$ of intervals together with the sets of considered change points, $\mathcal{T}(I), I \in \mathcal{I}$, are fixed. The parameters $\mathfrak{z}_k$ are then selected so that they provide the below prescribed features of the procedure under the parametric (time–homogeneous) model (3) with some fixed parameter vector $\boldsymbol{\theta}^*$.

Let $\hat{I}$ be the selected interval and $\hat{\boldsymbol{\theta}}$ be the corresponding adaptive estimate for data generated from a time–homogeneous parametric model. Both the interval $\hat{I}$ and estimate $\hat{\boldsymbol{\theta}}$ depends implicitly on the critical values $\mathfrak{z}_k$. Under the null hypothesis, the desirable feature of the adaptive procedure is that, with a high probability, it does not reject any interval I_k and selects the largest possible interval I_K. Equivalently, the selected interval $\hat{I}_k$ after the first k steps and the corresponding adaptive estimate $\hat{\boldsymbol{\theta}}_k$ should coincide with a high probability with their non–adaptive counterparts I_k and $\tilde{\boldsymbol{\theta}}_k = \tilde{\boldsymbol{\theta}}_{I_k}$.

Following Spokoiny and Chen (2007), this condition can be stated in the form

$$E_{\boldsymbol{\theta}^*} \left| 2\{L_{I_k}(\hat{\boldsymbol{\theta}}_k) - L_{I_k}(\tilde{\boldsymbol{\theta}}_k)\} \right|^r \leq \rho \mathfrak{R}_r(\boldsymbol{\theta}^*), \qquad k = 1, \ldots, K, \tag{9}$$

where ρ is a given positive constant and $\mathfrak{R}_r(\boldsymbol{\theta}^*)$ is the risk of the parametric estimation:

$$\mathfrak{R}_r(\boldsymbol{\theta}^*) = \max_{k \leq K} \mathrm{e}_{\boldsymbol{\theta}^*} \left| 2L_{I_k}\left(\tilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}^*\right) \right|^r.$$

In total, (9) states K conditions on the choice of K parameters $\mathfrak{z}_k$ that implicitly enter in the definition of the $\hat{\boldsymbol{\theta}}$'s. There are two ways to determine the values of $\mathfrak{z}_k$ by Monte Carlo simulations so that they satisfy (9). One possibility is to fix the values $\mathfrak{z}_k$ for each interval sequentially starting from $\mathfrak{z}_1$, $k = 1, \ldots, K$. An alternative way is to apply the critical values which linearly depend on $\log(|I_k|)$, $\mathfrak{z}_k = a + b \log(|I_K|/|I_k|)$, where $|I_k|$ denotes the length of interval I_k (Spokoiny and Chen (2007)).

3.2.3 Selecting the parameters r and ρ by minimizing the forecast error

The choice of critical values determined from (9) additionally depends on two "metaparameters" r and ρ. A simple strategy is to use a conservative values for these parameters and the corresponding set of critical values. On the other hand, the two parameters are global in the sense that they are independent of T. Hence, one can also determine them in a data–driven way by minimizing some global forecasting error. For instance, being interested in prediction at time T, we can compute the prediction of the conditional volatility in a period $T + h, h > 0$, using the (locally estimated) parametric model with the estimated parameters $\hat{\boldsymbol{\theta}}(T)$. Different values of r and ρ may lead to different estimates $\hat{\boldsymbol{\theta}}_{r,\rho}(T)$ and hence to different volatility forecasts $\hat{\sigma}^2_{r,\rho}(T + h|T)$. Following Cheng et al. (2003), the data driven choice of r and ρ can be done by minimizing the following objective function:

$$(\hat{r}, \hat{\rho}) = \operatorname*{argmin}_{r,\rho} \sum_T \sum_{h \in \mathcal{H}} \Lambda\big(R^2_{T+h}, \hat{\sigma}^2_{r,\rho}(T + h|T)\big), \tag{10}$$

where $\Lambda(\cdot, \cdot)$ is a loss function and $\mathcal{H}$ is the forecasting horizon set. For example, one can take $\Lambda(v, v') = |v - v'|$ or $\Lambda(v, v') = |v - v'|^2$. For daily data, the forecasting horizon could be one day, $\mathcal{H} = \{1\}$, or two weeks, $\mathcal{H} = \{1, \ldots, 10\}$.

4 Real–Data Application

In the real data application, we limit ourselves to the simplest conditional heteroscedasticity models: varying–coefficient constant volatility, ARCH(1), and the GARCH(1,1) models (for the sake of brevity, referred to also as the local constant, local ARCH, and local GARCH approximations). We first study the finite–sample critical values for the test of homogeneity by means of Monte Carlo simulations (Section 4.1). Later, we demonstrate the performance of the proposed pointwise adaptive estimation procedure discussed in Sections 3 for real data (Section 4.2). Throughout this section, we identify the GARCH(1,1) models by triplets (ω, α, β): for example, $(1, 0.1, 0.3)$–model. Simpler models, constant volatility and ARCH(1), are then indicated by $\alpha = \beta = 0$ and $\beta = 0$, respectively.

Table 1 Upper bounds on critical values $\mathfrak{z}(|I|)$ of the sequential supremum LR test defined by the intercept b_0 and slope b_1 of a line (11) for various parameter values of the ARCH(1) and GARCH(1,1) models; $r = 1, \rho = 1$.

Model (ω, α, β)	$\mathfrak{z}(10)$	Slope	$\mathfrak{z}(570)$
$(0.1, 0.0, 0.0)$	15.4	-0.55	5.5
$(0.1, 0.2, 0.0)$	16.6	-0.40	9.4
$(0.1, 0.4, 0.0)$	23.4	-0.74	10.1
$(0.1, 0.6, 0.0)$	30.8	-1.05	11.9
$(0.1, 0.8, 0.0)$	73.6	-3.37	16.4
$(0.1, 0.1, 0.8)$	19.5	-0.29	14.3
$(0.1, 0.2, 0.7)$	26.3	-0.68	14.1
$(0.1, 0.3, 0.6)$	25.1	-0.58	14.6
$(0.1, 0.4, 0.5)$	28.9	-0.74	15.6
$(0.1, 0.5, 0.4)$	29.8	-0.83	14.9
$(0.1, 0.6, 0.3)$	34.4	-1.05	15.5
$(0.1, 0.7, 0.2)$	27.1	-0.66	15.2
$(0.1, 0.8, 0.1)$	29.2	-0.75	15.7

4.1 Finite–sample critical values for the test of homogeneity

A practical application of the proposed adaptive procedure requires critical values for the test of local homogeneity of a time series. Since we rely on the supremum of the LR ratio test in a nonlinear model, and additionally, the test is applied sequentially in rather small samples, the only way to obtain a reasonable approximation of the critical values is to simulate them. Given an upper bound on the average risk between the adaptive and parametric

estimates under the null hypothesis of time homogeneity, which defined in (9) by two constants r and ρ, we determine upper bounds for critical values that are linearly proportional to the interval length:

$$\mathfrak{z}_k = b_0 + b_1 k = c_0 + c_1 \log(|I_K|/|I_k|) \tag{11}$$

for a given r and ρ. Since the critical values are generally decreasing with the interval length, the linear approximation cannot be used for an arbitrarily long interval. On the other hand, simulations show that the sequential nature of the search for the longest interval of homogeneity has only a small influence on the critical values: the critical values of independent supLR tests performed separately at various interval lengths and of a sequentially performed sequence of supLR tests are close to each other. Hence, we recommend to simulate critical values up to a certain interval length, e.g., $|I| = 500$, and to use the critical values obtained for the latest interval considered also for longer intervals if needed.

Unfortunately, the dependence on the parameters of the underlying model cannot be eliminated (in contrast to the case of local constant approximation). We simulated the critical values for ARCH(1) and GARCH(1,1) models with different values of underlying parameters; see Table 1 for critical values corresponding to $r = 1$ and $\rho = 1$. The adaptive estimation was performed sequentially on intervals with length ranging from $|I_0| = 10$ to $|I_K| = 570$ observations using a geometric grid with the initial interval length $m_0 = 10$ and multiplier $a = 1.25$, see Section 3.1. (Note however that the results are not sensitive to the choice of a.)

Generally, the critical values seem to increase with the values of the ARCH parameter or the sum of the ARCH and GARCH parameters. To deal with the dependence of the critical values on the underlying model parameters, we propose to choose the largest (most conservative) critical values corresponding to the any estimated parameter in the analyzed data. For example, if the largest estimated parameters of GARCH(1,1) are $\hat{\alpha} = 0.3$ and $\hat{\beta} = 0.8$, one should use $\mathfrak{z}(10) = 25.1$ and $\mathfrak{z}(570) = 14.6$. Note however that the proposed adaptive search procedure is not overly sensitive to this choice.

4.2 Stock index S&P 500

The proposed adaptive pointwise estimation method will be now applied to real time series consisting of the log–returns of the Standard & Poors 500 (S&P 500) stock index. To compare the local constant, local ARCH, and local GARCH with the standard parametric GARCH estimation[1] and predictions, we summarize the results concerning both parametric and adaptive methods

[1] The parametric GARCH is estimated at each time point using last two years of data (500 observations).

by looking at absolute prediction errors one–day ahead averaged over one month throughout this section, see (10) for $\Lambda(v, v') = |v - v'|$. Note that, to compute the prediction errors, we approximate the underlying volatility by squared returns. Despite being noisy, this approximation is unbiased and provides usually the correct ranking of methods (Andersen and Bollerslev (1998); Awartani and Corradi (2005)).

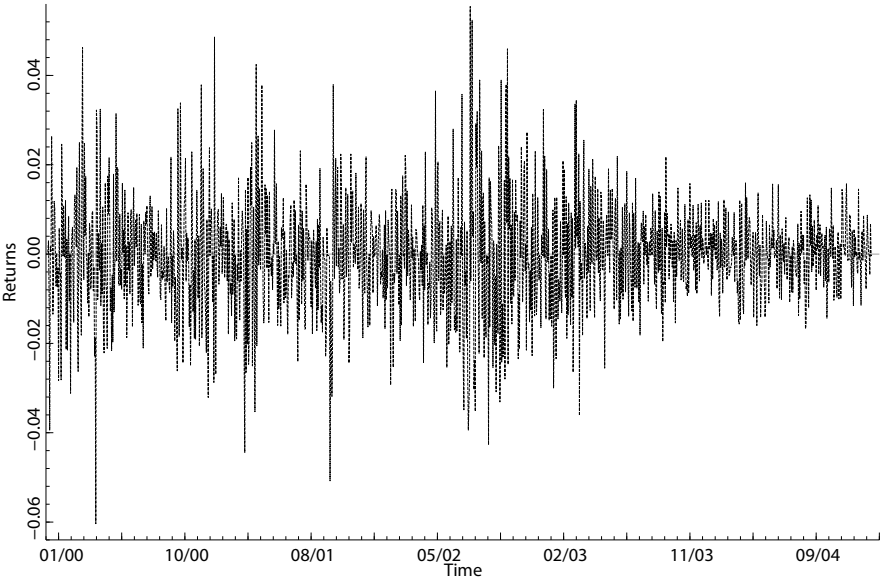

Fig. 1 The log–returns of S&P 500 from January 1, 2000 till December 31, 2004.

Now we concentrate on the S&P 500 stock index considered from January 1, 2000, to December 31, 2004, see Figure 1. This period is marked by many substantial events affecting the financial markets, ranging from September 11, 2001, terrorist attacks and the war in Iraq (2003) to the crash of the technology stock–market bubble (2000–2002). For the sake of simplicity, a particular time period is selected. The estimation results for years 2003 and 2004, where the first one represent a more volatile period (war on terrorism in Iraq) and the latter one is a less volatile period, are summarized on Fig. 2. It depicts the ratios of monthly prediction errors of all adaptive methods to the parametric GARCH ($r = 0.5$ and $\rho = 1.5$ for all methods).

In the beginning of 2003, which together with previous year 2002 corresponds to a more volatile period (see Figure 1), all adaptive methods detected rather quickly a structural break. Despite having just small amount of data after the break, all adaptive methods perform in the first half of 2003 as well as the parametric GARCH; the local constant approximation seems to be slightly better than other models though. In the middle of year 2003,

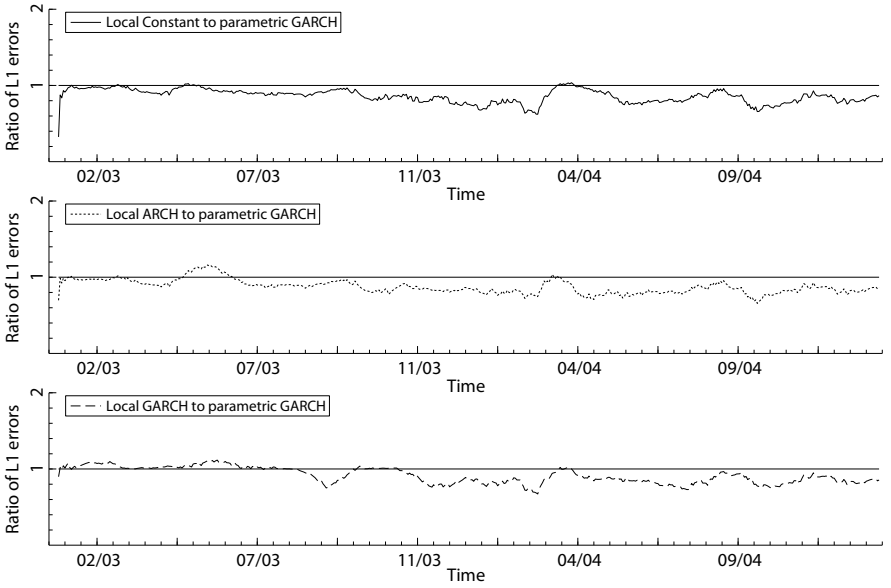

Fig. 2 The ratio of the absolute prediction errors of the three pointwise adaptive methods to the parametric GARCH for predictions one period ahead averaged over one month horizon. The S&P 500 index is considered from January, 2003 to December, 2004.

the local constant and local ARCH models are able to detect another structural change, possibly less pronounced than the one at the beginning of 2003 because of its late detection by the adaptive GARCH. Around this period, the local ARCH and local GARCH shortly performs slightly worse than the parametric GARCH. From the end of 2003 and in year 2004, all adaptive methods starts to outperform the parametric GARCH, where the reduction of the prediction errors due to the adaptive estimation amounts to 20% on average. Both local constant, local ARCH, and local GARCH methods exhibit a short period of instability in the first months of 2004, where their performance temporarily worsens to the level of parametric GARCH. This corresponds to "uncertainty" of the adaptive methods about the length of the interval of homogeneity. After this short period, the performance of all adaptive methods is comparable, although the local constant performs overall best of all methods (closely followed by local ARCH) judged by the global prediction error.

It seems that the benefit of pointwise adaptive estimation is most pronounced during periods of stability that follow an unstable period (i.e., in year 2004 here) rather than during a presumably rapidly changing environment. The reason is that, despite possible inconsistency of parametric methods under change points, the adaptive methods tend to have rather large variance when the intervals of time homogeneity become very short.

5 Conclusion

In this chapter, we extend the idea of adaptive pointwise estimation to more complex parametric models and demonstrate its use on the estimation of varying–coefficient conditional–heteroscedasticity models. The methodology is readily available for a wide class of conditional heteroscedasticity models, even though we concentrated on (G)ARCH models specifically. In the context of pointwise adaptive (G)ARCH estimation, we demonstrated that, on the one hand, the adaptive procedure, which itself depends on a number of auxiliary parameters, is rather insensitive to the choice of these parameters, and on the other hand, it facilitates the global selection of these parameters by means of fit or forecasting criteria.

References

Andersen, T. G. and Bollerslev, T. (1998): Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**, 885–905.

Andreou, E. and Ghysels, E. (2002): Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics* **17**, 579–600.

Andreou, E. and Ghysels, E. (2006): Monitoring disruptions in financial markets. *Journal of Econometrics* **135**, 77–124.

Andrews, D. W. K. (1993): Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**, 821–856.

Awartani, B. M. A. and Corradi, V. (2005): Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting* **21**, 167–183.

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996): Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **74**, 3–30.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Cai, Z., Fan, J. and Li, R. (2000): Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* **95**, 888–902.

Charles, A. and Darné, O. (2006): Large shocks and the September 11th terrorist attacks on international stock markets. *Economic Modelling* **23**, 683–698.

Chen, J. and Gupta, A. K. (1997): Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association* **92**, 739–747.

Cheng, M.-Y., Fan, J. and Spokoiny, V. (2003): Dynamic nonparametric filtering with application to volatility estimation. *In: Akritas, M. G., Politis, D. N. (Eds.): Recent Advances and Trends in Nonparametric Statistics*, 315–333. Elsevier, North Holland.

Chu, C. S. (1995): Detecting parameter shift in GARCH models. *Econometric Review* **14**, 241–266.

Diebold, F. X. and Inoue, A. (2001): Long memory and regime switching. *Journal of Econometrics* **105**, 131–159.

Ding, Z., Granger, C. W. J. and Engle, R. F. (1993): A long memory property of stock market returns and a new model. *Journal of Empirical Finance* **1**, 83–106.

Engle, R. F. (1982): Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1008.

Fan, J. and Zhang, W. (1999): Additive and varying coefficient models – statistical estimation in varying coefficient models. *The Annals of Statistics* **27**, 1491–1518.

Fan, J., Yao, Q. and Cai, Z. (2003): Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B* **65**, 57–80.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993): On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance* **48**, 1779–1801.

Hansen, B. and Lee, S.-W. (1994): Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory* **10**, 29–53.

Härdle, W., Herwatz, H. and Spokoiny, V. (2003): Time inhomogeneous multiple volatility modelling. *Journal of Financial Econometrics* **1**, 55–99.

Herwatz, H. and Reimers, H. E. (2001): Empirical modeling of the DEM/USD and DEM/JPY foreign exchange rate: structural shifts in GARCH-models and their implications. *SFB 373 Discussion Paper 2001/83* Humboldt-Universität zu Berlin, Germany.

Hillebrand, E. (2005): Neglecting parameter changes in GARCH models. *Journal of Econometrics* **129**, 121–138.

Kokoszka, P. and Leipus, R. (1999): Testing for Parameter Changes in ARCH Models. *Lithuanian Mathematical Journal* **39**, 231–247.

Kokoszka, P. and Leipus, R. (2000): Change-point estimation in ARCH models. *Bernoulli* **6**, 513–539.

Lin, S.-J. and Yang, J. (2000): Testing shifts in financial models with conditional heteroskedasticity: an empirical distribution function approach. *Research Paper 2000/30* University of Technology, Sydney, Australia.

Mercurio, D. and Spokoiny, V. (2004): Statistical inference for time-inhomogeneous volatility models. *The Annals of Statistics* **32**, 577–602.

Mikosch, T. and Stărică, C. (1999): Change of structure in financial time series, long range dependence and the GARCH model. *Manuscript, Department of Statistics*, University of Pennsylvania. See `http://www.math.chalmers.se/ starica/chp.pdf` or `http://citeseer.ist.psu.edu/mikosch99change.html`.

Mikosch, T. and Stărică, C. (2003): Non-stationarities in financial time series, the long-range dependence, and IGARCH effects. *The Review of Economics and Statistics* **86**, 378–390.

Mikosch, T. and Stărică, C. (2004): Changes of structure in financial time series and the GARCH model. *Revstat Statistical Journal* **2**, 41–73.

Nelson, D. B. (1990): ARCH models as diffusion approximations. *Journal of Econometrics* **45**, 7–38.

Nelson, D. B. (1991): Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**, 347–370.

Nocedal, J. and Wright, S. J. (1999): *Numerical optimization* Springer, Heidelberg.

Pesaran, M. H. and Timmermann, A. (2004): How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting* **20**, 411–425.

Polzehl, J. and Spokoiny, V. (2004): Varying coefficient GARCH versus local constant volatility modeling: comparison of the predictive power. *WIAS Discussion Paper 977* Berlin, Germany.

Polzehl, J. and Spokoiny, V. (2006): Propagation-separation approach for local likelihood estimation. *Probabability Theory and Related Fields* **135**, 335–362.

Sentana, E. (1995): Quadratic ARCH Models. *The Review of Economic Studies* **62**), 639–661.

Spokoiny, V. (1998): Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics* **26**, 1356–1378.

Spokoiny, V. and Chen, Y. (2007): Multiscale local change-point detection with with applications to Value-at-Risk. *The Annals of Statistics* to appear.

Stapf, J. and Werner, T. (2003): How wacky is DAX? The changing structure of German stock market volatility. *Discussion Paper 2003/18* Deutsche Bundesbank, Germany.

Taylor, S. J. (1986): *Modeling financial time series*. Wiley, Chichester.
Teyssiere, G. and Kokoszka, P. (2005): Change-point detection in GARCH models: asymptotic and bootstrap tests. *Journal of Business and Economic Statistics* to appear.

Extreme Value Theory for GARCH Processes

Richard A. Davis and Thomas Mikosch

Abstract We consider the extreme value theory for a stationary GARCH process with iid innovations. One of the basic ingredients of this theory is the fact that, under general conditions, GARCH processes have power law marginal tails and, more generally, regularly varying finite-dimensional distributions. Distributions with power law tails combined with weak dependence conditions imply that the scaled maxima of a GARCH process converge in distribution to a Fréchet distribution. The dependence structure of a GARCH process is responsible for the clustering of exceedances of a GARCH process above high and low level exceedances. The size of these clusters can be described by the extremal index. We also consider the convergence of the point processes of exceedances of a GARCH process toward a point process whose Laplace functional can be expressed explicitly in terms of the intensity measure of a Poisson process and a cluster distribution.

1 The Model

We consider a *GARCH process* $(X_t)_{t\in\mathbb{Z}}$ *of order* (p, q) (GARCH(p, q)) given by the equations[1]

Richard A. Davis
Department of Statistics, 1255 Amsterdam Avenue, Columbia University, New York, NY 10027, U.S.A., www.stat.columbia.edu/∼rdavis, e-mail: rdavis@stat.columbia.edu

Thomas Mikosch
Laboratory of Actuarial Mathematics, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark, www.math.ku.dk/∼mikosch, e-mail: mikosch@math.ku.dk

[1] Following the tradition in time series analysis, we index any stationary sequence (A_t) by the integers $\mathbb{Z}$. For practical purposes, one would consider the sequence $(X_t)_{t\in\mathbb{N}}$ corre-

DOI: 10.1007/978-3-540-71297-9_8,

$$X_t = \sigma_t Z_t\,, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2\,, \quad t \in \mathbb{Z}\,. \tag{1}$$

In order to avoid ambiguity as regards the order (p, q) we assume that $\alpha_p \beta_q \neq 0$ and that all coefficients α_i and β_j are non-negative. Since we are not interested in the trivial solution $X_t \equiv 0$ a.s. to (1) we also assume $\alpha_0 > 0$. Throughout, the *noise* or *innovations* sequence $(Z_t)_{t\in\mathbb{Z}}$ is an iid sequence with mean zero and unit variance.[2] We refer to (σ_t) as the *volatility* sequence of the GARCH process.

2 Strict Stationarity and Mixing Properties

To develop a reasonable extension of extreme value theory for iid sequences, the assumption of strict stationarity is often required. Therefore, we will always assume that there exists a unique strictly stationary causal[3] solution (X_t) to (1). Notice that stationarity[4] of the GARCH process (X_t) is equivalent to stationarity of the volatility sequence (σ_t). Necessary and sufficient conditions for the existence and uniqueness of a stationary ergodic solution to (1) are given in Nelson (1990) for the GARCH$(1, 1)$ case and for the general GARCH(p, q) case in Bougerol and Picard (1992); cf. Lindner (2008). These conditions will be discussed in Section 3.

Under general conditions such as the existence of a Lebesgue density of Z in some neighborhood of the origin, a stationary GARCH process (X_t) is *strongly mixing*, i.e.,

$$\sup_{C\in\mathcal{F}_{-\infty}^{0}, D\in\mathcal{F}_{t}^{\infty}} |P(C \cap D) - P(C)\,P(D)| = \alpha_t \to 0\,, \text{ as } t \to \infty\,,$$

where $\mathcal{F}_a^b$, $a \leq b$, is the σ-field generated by $(X_s)_{a\leq s\leq b}$ with the obvious modifications for infinite a or b. Moreover, the mixing rate α_t decays to 0 geometrically. These properties (under conditions on the distribution of Z more general than mentioned above) follow from work by Mokkadem (1990), cf. Doukhan (1994), Boussama (1998). Mokkadem (1990) and Boussama (1998) show that the process is in fact $\beta - mixing$. See Lindner (2008) for more details about the proof of strong mixing for GARCH processes.

---

sponding to observations at the times $t = 1, 2, \ldots$. If (A_t) is strictly stationary we write A for a generic element of the sequence.

[2] A standardization like unit variance of Z is necessary in order to avoid a trade-off in the scaling between σ_t and Z_t which would lead to non-identifiability of the parameters of the model (1).

[3] This means that X_t has representation as a measurable function of the past and present noise values Z_s, $s \leq t$.

[4] In what follows, we will use stationarity as a synonym for strict stationarity.

Substantial insight into the probabilistic structure of a GARCH process (X_t) is gained by embedding the squares X_t^2 and σ_t^2 into a stochastic recurrence equation. This procedure offers one a way to find conditions for stationarity of (X_t), but also for its marginal tail behavior and, in turn, for its extreme value behavior and the existence of moments.

3 Embedding a GARCH Process in a Stochastic Recurrence Equation

By the definition of a GARCH, the quantities X_t and σ_t are inextricably linked. Therefore any statement about the existence of the stationary sequence (X_t) and its distributional properties is also a statement about the corresponding properties of the stationary sequence (σ_t), and vice versa. Therefore one important approach to the understanding of the structure of a GARCH process is to express σ_t as a measurable function of past and present values Z_s, $s \leq t$. We refer to such a representation as a *causal solution* of (1).

Similarly to the representation of a causal ARMA process as a linear process of past values of the noise (see Brockwell and Davis (1991)) such a representation can be obtained by iterating the defining difference equation (1) for σ_t^2. Indeed, writing

$$\mathbf{Y}_t = \begin{pmatrix} \sigma_{t+1}^2 \\ \vdots \\ \sigma_{t-q+2}^2 \\ X_t^2 \\ \vdots \\ X_{t-p+1}^2 \end{pmatrix}, \quad \mathbf{A}_t = \begin{pmatrix} \alpha_1 Z_t^2 + \beta_1 & \beta_2 & \cdots & \beta_{q-1} & \beta_q & \alpha_2 & \alpha_3 & \cdots & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 \\ Z_t^2 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$\mathbf{B}_t = (\alpha_0, 0, \ldots, 0)',$$

we see that $((\mathbf{A}_t, \mathbf{B}_t))$ is an iid sequence, $\mathbf{Y}_{t-1}$ and $(\mathbf{A}_t, \mathbf{B}_t)$ are independent, where the $\mathbf{A}_t$'s are iid random $(p+q-1) \times (p+q-1)$ matrices and the $\mathbf{B}_t$'s iid $(p+q-1)$-dimensional random vectors. Then $(\mathbf{Y}_t)$ satisfies the following vector *stochastic recurrence equation* (SRE):

$$\mathbf{Y}_t = \mathbf{A}_t \mathbf{Y}_{t-1} + \mathbf{B}_t, \quad t \in \mathbb{Z}. \tag{2}$$

Iteration of the SRE (2) yields a unique stationary solution of the form

$$\mathbf{Y}_t = \mathbf{B}_t + \sum_{i=1}^{\infty} \mathbf{A}_t \cdots \mathbf{A}_{t-i+1} \mathbf{B}_{t-i}\,, \quad t \in \mathbb{Z}\,. \tag{3}$$

The crucial condition for the a.s. convergence of the infinite series in (3), hence for the existence of a strictly stationary solution to (2), is negativity of the *top Lyapunov exponent* given by

$$\gamma = \inf_{n \geq 1} n^{-1}\, E \log \|\mathbf{A}_n \cdots \mathbf{A}_1\|\,, \tag{4}$$

where $\|\cdot\|$ is the operator norm corresponding to a given norm in $\mathbb{R}^{p+q-1}$, see Bougerol and Picard (1992). By virtue of (3) the process $(\mathbf{Y}_t)$ has representation $\mathbf{Y}_t = f((Z_s)_{s \leq t})$ for some measurable function f. Therefore standard ergodic theory yields that $(\mathbf{Y}_t)$ is an ergodic process, see Krengel (1985).

In general, the top Lyapunov coefficient γ cannot be calculated explicitly, but a well known sufficient condition for $\gamma < 0$ is given by

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1\,, \tag{5}$$

see p. 122 in Bougerol and Picard (1992). Interestingly, (5) is also necessary and sufficient for the finiteness of $\mathrm{var}(X_t)$, hence for the second order stationarity of (X_t).

Example 1 (The GARCH(1,1) case)
This case is a real exception in the class of all GARCH processes since it is possible to calculate the Lyapunov coefficient explicitly. This is due to the fact that (2) essentially collapses into the one-dimensional SRE

$$\sigma_{t+1}^2 = \alpha_0 + (\alpha_1 Z_t^2 + \beta_1)\, \sigma_t^2\,, \tag{6}$$

where $A_t = \alpha_1 Z_t^2 + \beta_1$, hence

$$\gamma = n^{-1} E \log(A_n \cdots A_1) = E \log A = E \log(\alpha_1\, Z^2 + \beta_1)\,.$$

Then, following Nelson (1990), the conditions $E \log(\alpha_1\, Z^2 + \beta_1) < 0$ and $\alpha_0 > 0$ are necessary and sufficient for the existence of a stationary causal non-degenerate solution to (1). □

4 The Tails of a GARCH Process

The embedding of the squares X_t^2 and σ_t^2 of a stationary GARCH process (X_t) and its volatility sequence (σ_t) into the SRE (2) also allows one to use

classical theory about the tails of the solution to this SRE. Such a theory was developed by Kesten (1973), see also Goldie (1991) for an alternative approach, cf. Embrechts et al. (1997), Section 8.4.

Theorem 1 (Basrak et al. (2002))
Consider the process $(\mathbf{Y}_t)$ *in* (3) *obtained from embedding a stationary GARCH process into the SRE* (2). *Assume that* Z *has a positive density on* $\mathbb{R}$ *such that* $E(|Z|^h) < \infty$ *for* $h < h_0$ *and* $E(|Z|^{h_0}) = \infty$ *for some* $h_0 \in (0, \infty]$.[5]

Then there exist $\kappa > 0$ *and a finite-valued function* w *on the unit sphere*

$$\mathbb{S}^{p+q-2} = \{\mathbf{x} \in \mathbb{R}^{p+q-1} : |\mathbf{x}| = 1\}$$

for any fixed norm $|\cdot|$ *in* $\mathbb{R}^{p+q-1}$ *such that*

$$\text{for all } \mathbf{x} \in \mathbb{S}^{p+q-2}, \qquad \lim_{x\to\infty} x^{\kappa}\, P((\mathbf{x}, \mathbf{Y}) > x) = w(\mathbf{x}) \quad \text{exists}.$$

Moreover, if $\mathbf{x} \in \mathbb{S}^{p+q-2}$ *has non-negative components then* $w(\mathbf{x}) > 0$.

Furthermore, $\mathbf{Y}$ *is* regularly varying with index κ, [6] *i.e., there exist a constant* $c > 0$ *and a random vector* $\boldsymbol{\Theta}$ *on the unit sphere* $\mathbb{S}^{p+q-2}$ *such that for every* $t > 0$

$$x^{\kappa}\, P(|\mathbf{Y}| > t\,x\,,\, \mathbf{Y}/|\mathbf{Y}| \in \cdot\,) \xrightarrow{w} c\, t^{-\kappa} P(\boldsymbol{\Theta} \in \cdot\,), \quad \text{as } x \to \infty,$$

where $\xrightarrow{w}$ *denotes weak convergence on the Borel* σ*-field of* $\mathbb{S}^{p+q-2}$.

If we specify $\mathbf{x}$ to be a unit vector an immediate consequence is the following result.

Corollary 1 *Under the conditions on the distribution of* Z *in Theorem 1, the tails of the marginal distribution of a stationary GARCH process exhibit power law behavior: there exist* $\kappa > 0$ *and positive constants* $c_{|X|}$ *and* c_σ *such that*[7]

$$P(|X| > x) \sim c_{|X|}\, x^{-2\,\kappa} \quad \text{and} \quad P(\sigma > x) \sim c_\sigma\, x^{-2\,\kappa}.$$

With the exception of the GARCH$(1,1)$ case it is unknown how to evaluate c_σ, see Goldie (1991).

Applying a result by Breiman (1965), cf. Davis and Mikosch (2008), Section 4, one can also derive the relations[8]

[5] A condition such as $E(|Z|^{h_0}) = \infty$ is needed. It means that the distribution of Z is spread "sufficiently far out". Indeed, if Z is supported on too short an interval in the neighborhood of the origin, then $P(|X| > x)$ may decay to zero exponentially fast as $x \to \infty$, see Goldie and Grübel (1996).

[6] Basrak et al. (2002) proved this result under the condition that κ is not an even integer. Boman and Lindskog (2007) removed this condition.

[7] We write $f(x) \sim g(x)$ as $x \to \infty$ for two functions f and g whenever $f(x)/g(x) \to 1$ as $x \to \infty$.

[8] Here $x_\pm = \max(\pm x, 0)$ for $x \in \mathbb{R}$.

$$P(X > x) = P(\sigma Z_+ > x) \sim E((Z_+)^{2\kappa})\, P(\sigma > x)\,,$$
$$P(X \le -x) = P(-\sigma Z_- \le -x) \sim E((Z_-)^{2\kappa})\, P(\sigma > x)\,.$$

In the case of the general SRE model (2), the value of κ is determined as the solution to a complicated equation (see Kesten (1973)) which cannot be solved explicitly. As regards the general GARCH(p, q) model, one can estimate the *tail index* κ from observations of the GARCH process by using the tools of extreme value statistics, see Drees (2000) and Resnick and Stărică (1998), cf. Embrechts et al. (1997), Chapter 6, Resnick (2007). The GARCH$(1, 1)$ case again offers a more tractable form for κ.

Example 2 (The GARCH(1,1) case)
In this case, the SRE (2) collapses to the one-dimensional equation (6). Then direct calculation shows that Kesten's result yields κ as the unique solution (which exists under the conditions of Theorem 1) to the equation

$$E[A^\kappa] = E[(\alpha_1 Z^2 + \beta_1)^\kappa] = 1\,. \tag{7}$$

This equation can be solved for κ by numerical and/or simulation methods for fixed values of α_1 and β_1 from the stationarity region of a GARCH$(1, 1)$ process and assuming a concrete density for Z. In the case $\alpha_1 = 0.1$, Table 1 reports a small study for 2κ-values (the tail index of X and σ), assuming a standard normal or a unit variance student distribution with 4 degrees of freedom.

We mention that the inequality

$$E[(\alpha_1 Z^2 + \beta_1)^\kappa] \ge 1 \tag{8}$$

already implies that both $E(\sigma^{2\kappa})$ and $E(|X|^{2\kappa})$ are infinite. Indeed, since $\alpha_0 > 0$

$$E(\sigma_t^{2\kappa}) > E(\sigma_{t-1}^{2\kappa}\,(\alpha_1 Z_{t-1}^2 + \beta_1)^\kappa) = E(\sigma_t^{2\kappa})\, E((\alpha_1 Z^2 + \beta_1)^\kappa)\,.$$

By virtue of (8) this relation is impossible unless $E(\sigma^{2\kappa}) = \infty$. On the other hand, if

$$E[(\alpha_1 Z^2 + \beta_1)^\kappa] < 1 \tag{9}$$

a simple moment calculation involving Minkowski's inequality and representation (3) for $Y_t = \sigma_t^2$ imply that $E(|X|^{2\kappa}) = E(|Z|^{2\kappa})\, E(\sigma^{2\kappa}) < \infty$. Similar arguments for the existence of moments of the solution to a one-dimensional SRE can be found in Ling and McAleer (2002) and have been extended to GARCH and AGARCH models in Ling and McAleer (2002), cf. Lindner (2008). □

Example 3 (Integer valued κ and IGARCH)
Assume that we are in the GARCH$(1, 1)$ framework of Example 2. Since

Table 1 Results for 2κ when $\alpha_1 = 0.1$. Top: Standard normal noise. Bottom: Unit variance student noise with 4 degrees of freedom.

β_1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
2κ	2.0	12.5	16.2	18.5	20.2	21.7	23.0	24.2	25.4	26.5
β_1	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	
2κ	11.9	11.3	10.7	9.9	9.1	8.1	7.0	5.6	4.0	

β_1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
2κ	2.0	3.68	3.83	3.88	3.91	3.92	3.93	3.93	3.94	3.94
β_1	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	
2κ	3.65	3.61	3.56	3.49	3.41	3.29	3.13	2.90	2.54	

the solution κ of (7) is unique if it exists, it is possible to give necessary and sufficient conditions for tails of the form $P(|X| > x) \sim c_{|X|}\, x^{-2k}$ for $k = 1, 2, \ldots$. By virtue of the discussion of the consequences of the conditions (8) and (9) we may also conclude that $E(X^{2k}) < \infty$ if and only if $E[(\alpha_1 Z^2 + \beta_1)^k] < 1$.

Relation (7) turns into

$$1 = E[(\alpha_1 Z^2 + \beta_1)^k] = \beta_1^k \sum_{l=0}^{k} \binom{k}{l} (\alpha_1/\beta_1)^l \, E(Z^{2l}) \,. \tag{10}$$

For example, choosing $k = 1$ and recalling that $E(Z^2) = 1$, (10) turns into the equation $\alpha_1 + \beta_1 = 1$ which defines the integrated GARCH (IGARCH) model of Engle and Bollerslev (1986). It is a well known fact that a stationary IGARCH(p, q) process has infinite variance marginals.[9] In the IGARCH(1,1) case, Kesten's result yields the more sophisticated result $P(|X| > x) \sim c_{|X|}\, x^{-2}$. In applications to real-life data one often observes that the sum of the estimated parameters $\widehat{\alpha}_1 + \widehat{\beta}_1$ is close to 1 implying that moments slightly larger than two might not exist for a fitted GARCH process. An alternative explanation for this *IGARCH effect* are non-stationarities in the observed data as discussed in Mikosch and Stărică (2004).

For $k = 2$, (10) turns into the equation

$$1 = (\alpha_1 + \beta_1)^2 + \alpha_1^2 \, (E(Z^4) - 1) \,,$$

[9] This follows immediately by taking expectations in the defining equation (1) for σ_t^2 and recalling that $\alpha_0 > 0$.

implying that $P(|X| > x) \sim c_{|X|}\, x^{-4}$. Moreover, following the discussion above, the condition

$$1 > (\alpha_1 + \beta_1)^2 + \alpha_1^2\, (E(Z^4) - 1)$$

is necessary and sufficient for $E(X^4) < \infty$. □

5 Limit Theory for Extremes

5.1 Convergence of maxima

Generally, the asymptotic tail behavior of the marginal distribution of a stationary mixing sequence sequence (A_t) is the key factor in determining the weak limit behavior of the sequence of partial maxima

$$M_n(A) = \max_{i=1,\ldots,n} A_i\,, \quad n \geq 1\,. \tag{11}$$

This property is well documented in various monographs on the topic, see e.g., Leadbetter et al. (1983), Galambos (1987), Resnick (1987), Embrechts et al. (1997). In particular, for an iid sequence (A_t) with tail $P(A > x) \sim c\, x^{-\alpha}$ for some constant $c > 0$ it follows that

$$(c\, n)^{-1/\alpha} M_n(A) \xrightarrow{d} Y_\alpha\,, \tag{12}$$

where Y_α has the Fréchet distribution function,

$$\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \mathrm{e}^{\,-x^{-\alpha}}, & x > 0\,. \end{cases} \tag{13}$$

Now if (A_t) is a stationary sequence that satisfies a general mixing condition and the marginal distribution has the same tail behavior as above (i.e., $P(A > x) \sim c\, x^{-\alpha}$), then one often can show the existence of $\theta_A \in (0, 1]$ such that

$$(c\, n)^{-1/\alpha} M_n(A) \xrightarrow{d} \theta_A^{1/\alpha} Y_\alpha\,, \tag{14}$$

The parameter θ_A is called the *extremal index* and measures the level of clustering for extremes. The case $\theta_A = 1$ corresponds to no clustering in which case the limiting behavior of $M_n(A)$ is the same as for the maxima of an iid sequence. The reciprocal of the extremal index $1/\theta_A$ of a stationary sequence (A_t) has the interpretation as the expected size of clusters of high level exceedances in the sequence. For iid A_t's, $1/\theta_A = 1$, see (11), and there is no clustering of exceedances of high levels.

Returning to the GARCH setting, we assume that the conditions of Theorem 1 are satisfied. Then we know that $P(|X| > x) \sim c_{|X|}\, x^{-2\kappa}$ for some $\kappa, c_{|X|} > 0$, and we can even specify the value of κ in the GARCH$(1,1)$ case by solving the equation (7). The limit relation (12) is not directly applicable to the maxima $M_n(|X|)$ of the sequence $(|X_t|)$ because of the dependence in the sequence. However, it was mentioned in Section 2 that, under general conditions on the distribution of Z, the sequence (X_t), hence $(|X_t|)$ is strong mixing with geometric rate. Using this asymptotic independence condition it can be shown[10] that

$$(c_{|X|}\, n)^{-1/(2\,\kappa)}\, M_n(|X|) \xrightarrow{d} \theta_{|X|}^{1/(2\kappa)} Y_{2\,\kappa} \tag{15}$$

where $Y_{2\kappa}$ has the Fréchet distribution given in (13) and the extremal index $\theta_{|X|}$ is strictly less than one. In the aforementioned literature, it has also been shown that results analogous to (15) hold for the maxima of the sequences (σ_t) and (X_t) (at least when Z is symmetric) with the corresponding positive tail constants c_σ, c_X in (15) and extremal indices $\theta_\sigma, \theta_X \in (0,1)$ in (14).

Since the extremal index $\theta_{|X|}$ is strictly less than one for a GARCH process, the expected size of clusters of high level exceedances is $1/\theta_{|X|} > 1$. This is in sharp contrast to the case of stochastic volatility processes (see Davis and Mikosch (2008)), which have extremal indices that are equal to one and hence possess no clustering of extremes. Formulae for calculating the value of $\theta_{|X|}$ for a general GARCH(p,q) process are unknown, but in the ARCH(1) and GARCH$(1,1)$ cases more explicit expressions for $\theta_{|X|}$, θ_X and θ_σ exist. For example, in the GARCH$(1,1)$ case,

$$\theta_\sigma = \int_1^\infty P\left(\sup_{t\geq 1} \prod_{j=1}^{t} (\alpha_1\, Z_j^2 + \beta_1) \leq y^{-1}\right) \kappa\, y^{-\kappa-1}\, dy\,.$$

The right hand expression can be evaluated by Monte-Carlo simulations, see e.g. Haan et al. (1989) for the ARCH(1) case with standard normal noise Z, cf. Embrechts et al. (1997), Section 8.1, where one can also find some advice as to how the extremal index of a stationary sequence can be estimated from data.

5.2 Convergence of point processes

Advanced insight into the limit structure of the extremes of a stationary sequence (A_t) is provided by the weak convergence of the point processes

[10] See Haan et al. (1989) for ARCH(1); cf. Embrechts et al. (1997), Section 8.4, or Davis and Mikosch (1998); Mikosch and Stărică (2000) for GARCH(1,1), and Basrak et al. (2002) in the general GARCH(p,q) case.

$$N_n(\cdot) = \sum_{t=1}^{n} \varepsilon_{A_t/a_n}(\cdot)$$

toward a point process, whose points can be expressed in terms of products of Poisson points with independent points from a clustered distribution. Here ε_x denotes *Dirac measure* at x: for any set $B \subset \mathbb{R}$

$$\varepsilon_x(B) = \begin{cases} 1\,, & x \in B\,, \\ 0\,, & x \notin B\,, \end{cases}$$

and (a_n) is a suitably chosen sequence of positive constants. For example, for a GARCH process (X_t), (a_n) can be chosen such that

$$P(X > a_n) \sim n^{-1}\,.$$

Convergence in distribution, $N_n \xrightarrow{d} N$, of the sequence of point processes (N_n) toward the point process N is explained in standard books on point processes, see e.g. Kallenberg (1983), Daley and Vere–Jones (1988), Resnick (1987). Resnick's book describes the close relationship between the convergence of (N_n) and extreme value theory. For example, choosing the set $B = (x, \infty]$, $N_n \xrightarrow{d} N$ implies for the order statistics $A_{(1)} \le \cdots \le A_{(n)} = M_n(A)$ of the sample $A_1, \ldots, A_n$ that

$$\begin{aligned} P(N_n(x,\infty) < k) &= P(a_n^{-1} A_{(n-k+1)} \le x) \\ &\to P(N(x,\infty) < k) \\ &= \sum_{i=0}^{k-1} P(N(x,\infty) = i)\,, \quad x \in \mathbb{R}\,. \end{aligned}$$

Similar relations can be established for the joint convergence of finitely many order statistics in a sample, the joint convergence of the scaled minima $a_n^{-1}A_{(1)}$ and maxima $a_n^{-1}M_n(A)$, and various other results for extremes can be derived as well. Among others, the distribution of N determines the extremal index θ_A mentioned in Section 5.1.

The limit distribution of $(a_n^{-1}M_n(A))$ is determined by the distribution of the limiting point process N. For ARCH(1), GARCH(1, 1) and the general GARCH(p, q) processes (X_t) and their absolute values $(|X_t|)$ the form of the limit point process N was determined in Davis and Mikosch (1998), Mikosch and Stărică (2000) and Basrak et al. (2002), respectively. The Laplace functional of the limit point process N can be expressed explicitly in terms of the intensity measure of a Poisson process and the distribution of clusters. However, the general form of this representation has little practical value for probability computations.

5.3 The behavior of the sample autocovariance function

Basic measures of the dependence in a stationary time series (A_t) are the *autocovariance* and *autocorrelation functions* (ACVF and ACF) given respectively by

$$\gamma_A(h) = \mathrm{cov}(A_0, A_h) \quad \text{and} \quad \rho_A(h) = \mathrm{corr}(A_0, A_h) = \frac{\gamma_A(h)}{\gamma_A(0)}, \quad h \geq 0\,.$$

Here we have assumed that $\gamma_A(0) = \mathrm{var}(A) \in (0, \infty)$. Log-return series (X_t) usually have zero autocorrelations and therefore it is also common to consider the ACVFs and ACFs, $\gamma_{|X|^i}$ and $\rho_{|X|^i}$, $i = 1, 2$, of the absolute values $|X_t|$ and the squared returns X_t^2.

Since the ACVF and ACF of a stationary sequence (A_t) are in general unknown functions it is common to estimate them by their sample analogs (*sample ACVF* and *sample ACF*) given at lag $h \geq 0$ by

$$\widehat{\gamma}_A(h) = \frac{1}{n} \sum_{t=1}^{n-h} (A_t - \overline{A}_n)\,(A_{t+h} - \overline{A}_n)$$

and

$$\widehat{\rho}_A(h) = \frac{\widehat{\gamma}_A(h)}{\widehat{\gamma}_A(0)} = \frac{\sum_{t=1}^{n-h} (A_t - \overline{A}_n)\,(A_{t+h} - \overline{A}_n)}{\sum_{t=1}^{n} (A_t - \overline{A}_n)^2}\,,$$

where $\overline{A}_n = n^{-1} \sum_{t=1}^{n} A_t$ denotes the *sample mean*.

For non-linear processes (A_t) the limit theory for the sample ACVF and sample ACF is strongly influenced by heavy tails in the marginal distribution of A_t. This has been reported early on in Davis and Resnick (1996) for bilinear processes. In contrast, the sample ACF of a linear process (such as ARMA and FARIMA) consistently estimates the ACF of a Gaussian linear process with the same coefficients as the original linear process even when the distribution of A has such heavy tails that the variance of the process is infinite, in which case, the ACVF and the ACF are not defined. Interestingly, however, the rates of convergence in this heavy tailed case compare favorably to the usual $\sqrt{n}$-rates in the finite variance case, see Davis and Resnick (1985a, 1985b, 1986), cf. Brockwell and Davis (1991), Section 13.3.

The limit theory for the sample ACVF and ACF of a GARCH process (X_t), its absolute values and squares was studied in Davis and Mikosch (1998) in the ARCH(1) case, for GARCH(1,1) in Mikosch and Stărică (2000) and in Basrak et al. (2002) for the general GARCH(p, q) case as well as for solutions to SREs. The proofs of these results are based on the point process convergence results mentioned in Section 5.2, hence they are closely related to extreme value theory for GARCH processes. The limit distribution and the rates of convergence of the sample ACFs $\widehat{\rho}_X(h)$, $\widehat{\rho}_{|X|}(h)$ and $\widehat{\rho}_{X^2}(h)$ critically depend

on the tail index 2κ of the marginal distribution of a GARCH process, see Corollary 1. In particular, the following results hold.

1. If $2\kappa \in (0,2)$ then $\widehat{\rho}_X(h)$ and $\widehat{\rho}_{|X|}(h)$ have non-degenerate limit distributions. The same statement holds for $\widehat{\rho}_{X^2}(h)$ when $2\kappa \in (0,4)$.
2. If $2\kappa \in (2,4)$ then both $\widehat{\rho}_X(h)$, $\widehat{\rho}_{|X|}(h)$ converge in probability to their deterministic counterparts $\rho_X(h)$, $\rho_{|X|}(h)$, respectively, at the rate $n^{1-2/(2\kappa)}$ and the limit distribution, depending on an infinite variance stable distribution, is complex and difficult to describe.
3. If $2\kappa \in (4,8)$, then

$$n^{1-4/(4\kappa)}(\widehat{\rho}_{X^2}(h) - \rho_{X^2}(h)) \xrightarrow{d} S_\kappa(h)\,,$$

   where the random variable $S_\kappa(h)$ is a function of infinite variance stable random variables.
4. If $2\kappa > 4$ then the good mixing properties of the GARCH process (see Section 2) and standard central limit theory for stationary sequences (see e.g., Ibragimov and Linnik (1971) or Doukhan (1994)) imply that $(\widehat{\rho}_X(h))$ and $(\widehat{\rho}_{|X|}(h))$ have Gaussian limits at $\sqrt{n}$-rates. The corresponding result holds for (X_t^2) when $2\kappa > 8$.

These results show that the limit theory for the sample ACF of a GARCH process is rather complicated when low moments of the process do not exist. There is empirical evidence based on extreme value statistics indicating that log-return series might not have finite 4th or 5th moment,[11] and then the limit results above would show that the usual confidence bands for the sample ACF based on the central limit theorem and the corresponding $\sqrt{n}$-rates are far too optimistic in this case.

It is worth noting that for stochastic volatility models, an alternative class of models to the GARCH for modeling log-returns, the situation is markedly different. If the noise in the stochastic volatility process is chosen so that the marginal distribution matches the power law tail of the GARCH with index 2κ, then

$$(n/\ln n)^{1/(2\kappa)}\widehat{\rho}_X(h) \text{ and } (n/\ln n)^{1/(4\kappa)}\widehat{\rho}_{X^2}(h)$$

converge in distribution for $2\kappa \in (0,2)$ and $2\kappa \in (0,4)$, respectively. This illustrates the good large sample behavior of the sample ACF for stochastic volatility models even if ρ_X and ρ_{X^2} are not defined (see Davis and Mikosch (2001, 2008)) and illustrates another key difference between stochastic volatility and GARCH processes.

[11] See e.g. Embrechts et al. (1997), Chapter 6, and Mikosch (2003).

References

Basrak, B., Davis, R.A. and Mikosch, T. (2002): Regular variation of GARCH processes. *Stoch. Proc. Appl.* **99**, 95–116.

Boman, J. and Lindskog, F. (2007): Support theorems for the Radon transform and Cramér-Wold theorems. *Technical report, KTH Stockholm.*

Bougerol, P. and Picard, N. (1992): Stationarity of GARCH processes and of some non-negative time series. *J. Econometrics* **52**, 115–127.

Boussama, F. (1998): *Ergodicité, mélange et estimation dans le modelès GARCH.* PhD Thesis, Université 7 Paris.

Breiman, L. (1965): On some limit theorems similar to the arc-sine law. *Theory Probab. Appl.* **10**, 323–331.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods, 2nd edition.* Springer, Berlin, Heidelberg, New York.

Daley, D. and Vere–Jones, D. (1988): *An Introduction to the Theory of Point Processes.* Springer, Berlin.

Davis, R.A. and Mikosch, T. (1998): Limit theory for the sample ACF of stationary process with heavy tails with applications to ARCH. *Ann. Statist.* **26**, 2049–2080.

Davis, R.A. and Mikosch, T. (2001): The sample autocorrelations of financial time series models. *In: W.J. Fitzgerald, R.L. Smith, A.T. Walden and P.C. Young (Eds.) Nonlinear and Nonstationary Signal Processing*, 247–274. Cambridge University Press, Cambridge.

Davis, R.A. and Mikosch, T. (2008): Probabilistic properties of stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 255–267. Springer, New York.

Davis, R.A. and Resnick, S.I. (1985a): Limit theory for moving averages of random variables with regularly varying tail probabilities. *Ann. Probab.* **13**, 179–195.

Davis, R.A. and Resnick, S.I. (1985b): More limit theory for the sample correlation function of moving averages. *Stoch. Proc. Appl.* **20**, 257–279.

Davis, R.A. and Resnick, S.I. (1986): Limit theory for the sample covariance and correlation functions of moving averages. *Ann. Statist.* **14**, 533–558.

Davis, R.A. and Resnick, S.I. (1996): Limit theory for bilinear processes with heavy tailed noise. *Ann. Appl. Probab.* **6**, 1191–1210.

Doukhan, P. (1994): *Mixing. Properties and Examples. Lecture Notes in Statistics* **85**, Springer, New York.

Drees, H. (2000): Weighted approximations of tail processes for β-mixing random variables. *Ann. Appl. Probab.* **10**, 1274–1301.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997): *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

Engle, R.F. and Bollerslev, T. (1986): Modelling the persistence of conditional variances. With comments and a reply by the authors. *Econometric Rev.* **5**, 1–87.

Galambos, J. (1987): *Asymptotic Theory of Extreme Order Statistics* (2nd edition). Krieger, Malabar, Florida.

Goldie, C.M. (1991): Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.* **1**, 126–166.

Goldie, C.M. and Grübel, R. (1996): Perpetuities with thin tails. *Adv. Appl. Probab.* **28**, 463–480.

Haan, L. de, Resnick, S.I., Rootzén, H. and Vries, C.G. de (1989): Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. *Stoch. Proc. Appl.* **32**, 213–224.

Hult, H. and Lindskog, F. (2006): On Kesten's counterexample to the Cramér-Wold device for regular variation. *Bernoulli* **12**, 133–142.

Ibragimov, I.A. and Linnik, Yu.V. (1971): *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

Kallenberg, O. (1983): *Random Measures, 3rd edition.* Akademie–Verlag, Berlin.

Kesten, H. (1973): Random difference equations and renewal theory for products of random matrices. *Acta Math.* **131**, 207–248.

Krengel, U. (1985): *Ergodic Theorems.* De Gruyter, Berlin.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983): *Extremes and Related Properties of Random Sequences and Processes.* Springer, Berlin.

Lindner, A.M. (2008): Stationarity, Mixing, Distributional Properties and Moments of GARCH(p, q)–Processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 43–69. Springer, New York.

Ling, S. and McAleer, M. (2002): Stationarity and the existence of moments of a family of GARCH processes. *J. Econometrics* **106**, 109–117.

Ling, S. and McAleer, M. (2002): Necessary and sufficient conditions for the GARCH(r, s) and asymmetric power GARCH(r, s) models. *Econometric Theory* **18**, 722–729.

Mikosch, T. (2003): Modelling dependence and tails of financial time series. *In: B. Finkenstädt and H. Rootzén (Eds.): Extreme Values in Finance, Telecommunications and the Environment*, 185–286. Chapman and Hall.

Mikosch, T. and Stărică, C. (2000): Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Ann. Stat.* **28**, 1427–1451.

Mikosch, T. and Stărică, C. (2004): Non-stationarities in financial time series, the long-range dependence and the IGARCH effects. *Review of Economics and Statistics* **86**, 378–390.

Mokkadem, A. (1990): Propriétés de mélange des processus autoregréssifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist.* **26**, 219–260.

Nelson, D.B. (1990): Stationarity and persistence in the GARCH$(1, 1)$ model. *Econometric Theory* **6**, 318–334.

Resnick, S.I. (1987): *Extreme Values, Regular Variation, and Point Processes.* Springer, New York.

Resnick, S.I. (2007): *Heavy Tail Phenomena; Probabilistic and Statistical Modeling.* Springer, New York.

Resnick, S.I. and Stărică, C. (1998): Tail index estimation for dependent data. *Ann. Appl. Probab.* **8**, 1156–1183.

Multivariate GARCH Models

Annastiina Silvennoinen and Timo Teräsvirta*

Abstract This article contains a review of multivariate GARCH models. Most common GARCH models are presented and their properties considered. This also includes nonparametric and semiparametric models. Existing specification and misspecification tests are discussed. Finally, there is an empirical example in which several multivariate GARCH models are fitted to the same data set and the results compared.

1 Introduction

Modelling volatility in financial time series has been the object of much attention ever since the introduction of the Autoregressive Conditional Heteroskedasticity (ARCH) model in the seminal paper of Engle (1982). Subsequently, numerous variants and extensions of ARCH models have been proposed. A large body of this literature has been devoted to univariate models;

Annastiina Silvennoinen
School of Finance and Economics, University of Technology Sydney, Box 123, Broadway NSW 2007, e-mail: annastiina.silvennoinen@uts.edu.au

Timo Teräsvirta
CREATES, School of Economics and Management, University of Aarhus, DK-8000 Aarhus C, and Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, e-mail: tterasvirta@econ.au.dk

* This research has been supported by the Jan Wallander and Tom Hedelius Foundation, Grant No. P2005–0033:1, and the Danish National Research Foundation. We thank Robert Engle for providing the data for the empirical section, Markku Lanne and Pentti Saikkonen for sharing with us their program code for their Generalized Orthogonal Factor GARCH model, and Mika Meitz for programming assistance. Part of the research was done while the first author was visiting CREATES, University of Aarhus, whose kind hospitality is gratefully acknowledged. The responsibility for any errors and shortcomings in this paper remains ours.

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_9,

see for example Bollerslev et al. (1994), Palm (1996), Shephard (1996), and several chapters of this Handbook for surveys of this literature.

While modelling volatility of the returns has been the main centre of attention, understanding the comovements of financial returns is of great practical importance. It is therefore important to extend the considerations to multivariate GARCH (MGARCH) models. For example, asset pricing depends on the covariance of the assets in a portfolio, and risk management and asset allocation relate for instance to finding and updating optimal hedging positions. For examples, see Bollerslev et al. (1988), Ng (1991), and Hansson and Hordahl (1998). Multivariate GARCH models have also been used to investigate volatility and correlation transmission and spillover effects in studies of contagion, see Tse and Tsui (2002) and Bae et al. (2003).

What then should the specification of an MGARCH model be like? On one hand, it should be flexible enough to be able to represent the dynamics of the conditional variances and covariances. On the other hand, as the number of parameters in an MGARCH model often increases rapidly with the dimension of the model, the specification should be parsimonious enough to enable relatively easy estimation of the model and also allow for easy interpretation of the model parameters. However, parsimony often means simplification, and models with only a few parameters may not be able to capture the relevant dynamics in the covariance structure. Another feature that needs to be taken into account in the specification is imposing positive definiteness (as covariance matrices need, by definition, to be positive definite). One possibility is to derive conditions under which the conditional covariance matrices implied by the model are positive definite, but this is often infeasible in practice. An alternative is to formulate the model in a way that positive definiteness is implied by the model structure (in addition to some simple constraints).

Combining these needs has been the difficulty in the MGARCH literature. The first GARCH model for the conditional covariance matrices was the so-called VEC model of Bollerslev et al. (1988), see Engle et al. (1984) for an ARCH version. This model is a very general one, and a goal of the subsequent literature has been to formulate more parsimonious models. Furthermore, since imposing positive definiteness of the conditional covariance matrix in this model is difficult, formulating models with this feature has been considered important. Furthermore, constructing models in which the estimated parameters have direct interpretation has been viewed as beneficial.

In this paper, we survey the main developments of the MGARCH literature. For another such survey, see Bauwens et al. (2006). This paper is organized as follows. In Section 2, several MGARCH specifications are reviewed. Statistical properties of the models are the topic of Section 3, whereas testing MGARCH models is discussed in Section 4. An empirical comparison of a selection of the models is given in Section 5. Finally, some conclusions and directions for future research are provided in Section 6.

2 Models

Consider a stochastic vector process $\{\boldsymbol{r}_t\}$ with dimension $N \times 1$ such that $E\boldsymbol{r}_t = \boldsymbol{0}$. Let $\mathcal{F}_{t-1}$ denote the information set generated by the observed series $\{\boldsymbol{r}_t\}$ up to and including time $t-1$. We assume that $\boldsymbol{r}_t$ is conditionally heteroskedastic:

$$\boldsymbol{r}_t = \boldsymbol{H}_t^{1/2}\boldsymbol{\eta}_t \tag{1}$$

given the information set $\mathcal{F}_{t-1}$, where the $N \times N$ matrix $\boldsymbol{H}_t = [h_{ijt}]$ is the conditional covariance matrix of $\boldsymbol{r}_t$ and $\boldsymbol{\eta}_t$ is an iid vector error process such that $E\boldsymbol{\eta}_t\boldsymbol{\eta}_t' = \boldsymbol{I}$. This defines the standard multivariate GARCH framework, in which there is no linear dependence structure in $\{\boldsymbol{r}_t\}$. In financial applications, $\boldsymbol{r}_t$ is most often viewed as a vector of log-returns of N assets.

What remains to be specified is the matrix process $\boldsymbol{H}_t$. Various parametric formulations will be reviewed in the following subsections. We have divided these models into four categories. In the first one, the conditional covariance matrix $\boldsymbol{H}_t$ is modelled directly. This class includes, in particular, the VEC and BEKK models to be defined in Section 2.1 that were among the first parametric MGARCH models. The models in the second class, the factor models, are motivated by parsimony: the process $\boldsymbol{r}_t$ is assumed to be generated by a (small) number of unobserved heteroskedastic factors. Models in the third class are built on the idea of modelling the conditional variances and correlations instead of straightforward modelling of the conditional covariance matrix. Members of this class include the Constant Conditional Correlation (CCC) model and its extensions. The appeal of this class lies in the intuitive interpretation of the correlations, and models belonging to it have received plenty of attention in the recent literature. Finally, we consider semi- and nonparametric approaches that can offset the loss of efficiency of the parametric estimators due to misspecified structure of the conditional covariance matrices. Multivariate stochastic volatility models are discussed in a separate chapter of this Handbook, see Chib et al. (2008).

Before turning to the models, we discuss some points that need attention when specifying an MGARCH model. As already mentioned, a problem with MGARCH models is that the number of parameters can increase very rapidly as the dimension of $\boldsymbol{r}_t$ increases. This creates difficulties in the estimation of the models, and therefore an important goal in constructing new MGARCH models is to make them reasonably parsimonious while maintaining flexibility. Another aspect that has to be imposed is the positive definiteness of the conditional covariance matrices. Ensuring positive definiteness of a matrix, usually through an eigenvalue-eigenvector-decomposition, is a numerically difficult problem, especially in large systems. Yet another difficulty with MGARCH models has to do with the numerical optimization of the likelihood function (in the case of parametric models). The conditional covariance (or correlation) matrix appearing in the likelihood depends on the time index t, and often has to be inverted for all t in every iteration of the numerical

optimization. When the dimension of $\boldsymbol{r}_t$ increases, this is both a time consuming and numerically unstable procedure. Avoiding excessive inversion of matrices is thus a worthy goal in designing MGARCH models. It should be emphasized, however, that practical implementation of all the models to be considered in this chapter is of course feasible, but the problem lies in devising easy to use, automated estimation routines that would make widespread use of these models possible.

2.1 Models of the conditional covariance matrix

The VEC-GARCH model of Bollerslev et al. (1988) is a straightforward generalization of the univariate GARCH model. Every conditional variance and covariance is a function of all lagged conditional variances and covariances, as well as lagged squared returns and cross-products of returns. The model may be written as follows:

$$\mathrm{vech}(\boldsymbol{H}_t) = \boldsymbol{c} + \sum_{j=1}^{q} \boldsymbol{A}_j \mathrm{vech}(\boldsymbol{r}_{t-j}\boldsymbol{r}'_{t-j}) + \sum_{j=1}^{p} \boldsymbol{B}_j \mathrm{vech}(\boldsymbol{H}_{t-j})\,, \qquad (2)$$

where $\mathrm{vech}(\cdot)$ is an operator that stacks the columns of the lower triangular part of its argument square matrix, $\boldsymbol{c}$ is an $N(N+1)/2 \times 1$ vector, and $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ are $N(N+1)/2 \times N(N+1)/2$ parameter matrices. In fact, the authors introduced a multivariate GARCH–in-mean model, but in this chapter we only consider its conditional covariance component. The generality of the VEC model is an advantage in the sense that the model is very flexible, but it also brings disadvantages. One is that there exist only sufficient, rather restrictive, conditions for $\boldsymbol{H}_t$ to be positive definite for all t, see Gouriéroux (1997), Chapter 6. Besides, the number of parameters equals $(p+q)(N(N+1)/2)^2 + N(N+1)/2$, which is large unless N is small. Furthermore, as will be discussed below, estimation of the parameters is computationally demanding.

Bollerslev et al. (1988) presented a simplified version of the model by assuming that $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ in (2) are diagonal matrices. In this case, it is possible to obtain conditions for $\boldsymbol{H}_t$ to be positive definite for all t, see Bollerslev et al. (1994). Estimation is less difficult than in the complete VEC model because each equation can be estimated separately. But then, this 'diagonal VEC' model that contains $(p+q+1)N(N+1)/2$ parameters seems too restrictive since no interaction is allowed between the different conditional variances and covariances.

A numerical problem is that estimation of parameters of the VEC model is computationally demanding. Assuming that the errors $\boldsymbol{\eta}_t$ follow a multivariate normal distribution, the log-likelihood of the model (1) has the following form:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) = c - (1/2)\sum_{t=1}^{T} \ln|\boldsymbol{H}_t| - (1/2)\sum_{t=1}^{T} \boldsymbol{r}_t' \boldsymbol{H}_t^{-1} \boldsymbol{r}_t. \tag{3}$$

The parameter vector $\boldsymbol{\theta}$ has to be estimated iteratively. It is seen from (3) that the conditional covariance matrix $\boldsymbol{H}_t$ has to be inverted for every t in each iteration, which may be tedious when the number of observations is large and when, at the same time, N is not small. An even more difficult problem, is how to ensure positive definiteness of the covariance matrices. In the case of the VEC model there does not seem to exist a general solution to this problem. The problem of finding the necessary starting-values for $\boldsymbol{H}_t$ is typically solved by using the estimated unconditional covariance matrix as the initial value.

A model that can be viewed as a restricted version of the VEC model is the Baba-Engle-Kraft-Kroner (BEKK) model defined in Engle and Kroner (1995). It has the attractive property that the conditional covariance matrices are positive definite by construction. The model has the form

$$\boldsymbol{H}_t = \boldsymbol{C}\boldsymbol{C}' + \sum_{j=1}^{q}\sum_{k=1}^{K} \boldsymbol{A}_{kj}' \boldsymbol{r}_{t-j} \boldsymbol{r}_{t-j}' \boldsymbol{A}_{kj} + \sum_{j=1}^{p}\sum_{k=1}^{K} \boldsymbol{B}_{kj}' \boldsymbol{H}_{t-j} \boldsymbol{B}_{kj}\,, \tag{4}$$

where $\boldsymbol{A}_{kj}$, $\boldsymbol{B}_{kj}$, and $\boldsymbol{C}$ are $N \times N$ parameter matrices, and $\boldsymbol{C}$ is lower triangular. The decomposition of the constant term into a product of two triangular matrices is to ensure positive definiteness of $\boldsymbol{H}_t$. The BEKK model is covariance stationary if and only if the eigenvalues of $\sum_{j=1}^{q}\sum_{k=1}^{K} \boldsymbol{A}_{kj} \otimes \boldsymbol{A}_{kj} + \sum_{j=1}^{p}\sum_{k=1}^{K} \boldsymbol{B}_{kj} \otimes \boldsymbol{B}_{kj}$, where $\otimes$ denotes the Kronecker product of two matrices, are less than one in modulus. Whenever $K > 1$ an identification problem arises because there are several parameterizations that yield the same representation of the model. Engle and Kroner (1995) give conditions for eliminating redundant, observationally equivalent representations.

Interpretation of parameters of (4) is not easy. But then, consider the first order model

$$\boldsymbol{H}_t = \boldsymbol{C}\boldsymbol{C}' + \boldsymbol{A}'\boldsymbol{r}_{t-1}\boldsymbol{r}_{t-1}'\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{H}_{t-1}\boldsymbol{B}. \tag{5}$$

Setting $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix, (5) becomes

$$\boldsymbol{H}_t = \boldsymbol{C}\boldsymbol{C}' + \boldsymbol{A}'\boldsymbol{r}_{t-1}\boldsymbol{r}_{t-1}'\boldsymbol{A} + \boldsymbol{D}E[\boldsymbol{A}'\boldsymbol{r}_{t-1}\boldsymbol{r}_{t-1}'\boldsymbol{A}|\mathcal{F}_{t-2}]\boldsymbol{D}. \tag{6}$$

It is seen from (6) that what is now modelled are the conditional variances and covariances of certain linear combinations of the vector of asset returns $\boldsymbol{r}_t$ or 'portfolios'. Kroner and Ng (1998) restrict $\boldsymbol{B} = \delta\boldsymbol{A}$, where $\delta > 0$ is a scalar.

A further simplified version of (5) in which $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal matrices has sometimes appeared in applications. This 'diagonal BEKK' model trivially satisfies the equation $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{D}$. It is a restricted version of the diagonal VEC model such that the parameters of the covariance equations (equations

for h_{ijt}, $i \neq j$) are products of the parameters of the variance equations (equations for h_{iit}). In order to obtain a more general model (that is, to relax these restrictions on the coefficients of the covariance terms) one has to allow $K > 1$. The most restricted version of the diagonal BEKK model is the scalar BEKK one with $\boldsymbol{A} = a\boldsymbol{I}$ and $\boldsymbol{B} = b\boldsymbol{I}$, where a and b are scalars.

Each of the BEKK models implies a unique VEC model, which then generates positive definite conditional covariance matrices. Engle and Kroner (1995) provide sufficient conditions for the two models, BEKK and VEC, to be equivalent. They also give a representation theorem that establishes the equivalence of diagonal VEC models (that have positive definite covariance matrices) and general diagonal BEKK models. When the number of parameters in the BEKK model is less than the corresponding number in the VEC model, the BEKK parameterization imposes restrictions that makes the model different from that of VEC model. Increasing K in (4) eliminates those restrictions and thus increases the generality of the BEKK model towards the one obtained from using pure VEC model. Engle and Kroner (1995) give necessary conditions under which all unnecessary restrictions are eliminated. However, too large a value of K will give rise to the identification problem mentioned earlier.

Estimation of a BEKK model still involves somewhat heavy computations due to several matrix inversions. The number of parameters, $(p+q)KN^2 + N(N+1)/2$ in the full BEKK model, or $(p+q)KN+N(N+1)/2$ in the diagonal one, is still quite large. Obtaining convergence may therefore be difficult because (4) is not linear in parameters. There is the advantage, however, that the structure automatically ensures positive definiteness of $\boldsymbol{H}_t$, so this does not need to be imposed separately. Partly because numerical difficulties are so common in the estimation of BEKK models, it is typically assumed $p = q = K = 1$ in applications of (4).

Parameter restrictions to ensure positive definiteness are not needed in the matrix exponential GARCH model proposed by Kawakatsu (2006). It is a generalization of the univariate exponential GARCH model of Nelson (1991) and is defined as follows:

$$\begin{aligned}\operatorname{vech}(\ln \boldsymbol{H}_t - \boldsymbol{C}) = &\sum_{i=i}^{q} \boldsymbol{A}_i \boldsymbol{\eta}_{t-i} + \sum_{i=1}^{q} \boldsymbol{F}_i(|\boldsymbol{\eta}_{t-i}| - E|\boldsymbol{\eta}_{t-i}|) \\ &+ \sum_{i=1}^{p} \boldsymbol{B}_i \operatorname{vech}(\ln \boldsymbol{H}_{t-i} - \boldsymbol{C}), \end{aligned} \tag{7}$$

where $\boldsymbol{C}$ is a symmetric $N \times N$ matrix, and $\boldsymbol{A}_i$, $\boldsymbol{B}_i$, and $\boldsymbol{F}_i$ are parameter matrices of sizes $N(N+1)/2 \times N$, $N(N+1)/2 \times N(N+1)/2$, and $N(N+1)/2 \times N$, respectively. There is no need to impose restrictions on the parameters to ensure positive definiteness, because the matrix $\ln \boldsymbol{H}_t$ need not be positive definite. The positive definiteness of the covariance matrix $\boldsymbol{H}_t$ follows from the fact that for any symmetric matrix $\boldsymbol{S}$, the matrix exponential defined as

$$\exp(\boldsymbol{S}) = \sum_{i=0}^{\infty} \frac{\boldsymbol{S}^i}{i!}$$

is positive definite. Since the model contains a large number of parameters, Kawakatsu (2006) discusses a number of more parsimonious specifications. He also considers the estimation of the model, hypothesis testing, the interpretation of the parameters, and provides an application. How popular this model will turn out in practice remains to be seen.

2.2 Factor models

Factor models are motivated by economic theory. For instance, in the arbitrage pricing theory of Ross (1976) returns are generated by a number of common unobserved components, or factors; for further discussion see Engle et al. (1990) who introduced the first factor GARCH model. In this model it is assumed that the observations are generated by underlying factors that are conditionally heteroskedastic and possess a GARCH-type structure. This approach has the advantage that it reduces the dimensionality of the problem when the number of factors relative to the dimension of the return vector $\boldsymbol{r}_t$ is small.

Engle et al. (1990) define a factor structure for the conditional covariance matrix as follows. They assume that $\boldsymbol{H}_t$ is generated by K ($< N$) underlying, not necessarily uncorrelated, factors $f_{k,t}$ as follows:

$$\boldsymbol{H}_t = \boldsymbol{\Omega} + \sum_{k=1}^{K} \boldsymbol{w}_k \boldsymbol{w}_k' f_{k,t}\,, \tag{8}$$

where $\boldsymbol{\Omega}$ is an $N \times N$ positive semi-definite matrix, $\boldsymbol{w}_k$, $k = 1, \ldots, K$, are linearly independent $N \times 1$ vectors of factor weights, and the $f_{k,t}$'s are the factors. It is assumed that these factors have a first-order GARCH structure:

$$f_{k,t} = \omega_k + \alpha_k(\boldsymbol{\gamma}_k' \boldsymbol{r}_{t-1})^2 + \beta_k f_{k,t-1}\,,$$

where ω_k, α_k, and β_k are scalars and $\boldsymbol{\gamma}_k$ is an $N \times 1$ vector of weights. The number of factors K is intended to be much smaller than the number of assets N, which makes the model feasible even for a large number of assets. Consistent but not efficient two-step estimation method using maximum likelihood is discussed in Engle et al. (1990). In their application, the authors consider two factor-representing portfolios as the underlying factors that drive the volatilities of excess returns of the individual assets. One factor consists of value-weighted stock index returns and the other one of average T-bill returns of different maturities. This choice is motivated by principal component analysis.

Diebold and Nerlove (1989) propose a model similar to the one formulated in Engle et al. (1990). However their model is rather a stochastic volatility model than a GARCH one, and hence we do not discuss its properties here; see Sentana (1998) for a comparison of this model with the factor GARCH one.

In the factor ARCH model of Engle et al. (1990) the factors are generally correlated. This may be undesirable as it may turn out that several of the factors capture very similar characteristics of the data. If the factors were uncorrelated, they would represent genuinely different common components driving the returns. Motivated by this consideration, several factor models with uncorrelated factors have been proposed in the literature. In all of them, the original observed series contained in $\boldsymbol{r}_t$ are assumed to be linked to unobserved, uncorrelated variables, or factors, $\boldsymbol{z}_t$ through a linear, invertible transformation $\boldsymbol{W}$:

$$\boldsymbol{r}_t = \boldsymbol{W}\boldsymbol{z}_t\,,$$

where $\boldsymbol{W}$ is thus a nonsingular $N \times N$ matrix. Use of uncorrelated factors can potentially reduce their number relative to the approach where the factors can be correlated. The unobservable factors are estimated from the data through $\boldsymbol{W}$. The factors $\boldsymbol{z}_t$ are typically assumed to follow a GARCH process. Differences between the factor models are due to the specification of the transformation $\boldsymbol{W}$ and, importantly, whether the number of heteroskedastic factors is less than the number of assets or not.

In the Generalized Orthogonal (GO–) GARCH model of van der Weide (2002), the uncorrelated factors $\boldsymbol{z}_t$ are standardized to have unit unconditional variances, that is, $E\boldsymbol{z}_t\boldsymbol{z}_t' = \boldsymbol{I}$. This specification extends the Orthogonal (O–) GARCH model of Alexander and Chibumba (1997) in that $\boldsymbol{W}$ is not required to be orthogonal, only invertible. The factors are conditionally heteroskedastic with GARCH-type dynamics. The $N \times N$ diagonal matrix of conditional variances of $\boldsymbol{z}_t$ is defined as follows:

$$\boldsymbol{H}_t^z = (\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{B}) + \boldsymbol{A} \odot (\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}') + \boldsymbol{B}\boldsymbol{H}_{t-1}^z\,, \tag{9}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal $N \times N$ parameter matrices and $\odot$ denotes the Hadamard (i.e. elementwise) product of two conformable matrices. The form of the constant term imposes the restriction $E\boldsymbol{z}_t\boldsymbol{z}_t' = \boldsymbol{I}$. Covariance stationarity of $\boldsymbol{r}_t$ in the models with uncorrelated factors is ensured if the diagonal elements of $\boldsymbol{A} + \boldsymbol{B}$ are less than one. Therefore the conditional covariance matrix of $\boldsymbol{r}_t$ can be expressed as

$$\boldsymbol{H}_t = \boldsymbol{W}\boldsymbol{H}_t^z\boldsymbol{W}' = \sum_{k=1}^{N} \boldsymbol{w}_{(k)}\boldsymbol{w}_{(k)}' h_{k,t}^z\,, \tag{10}$$

where $\boldsymbol{w}_{(k)}$ are the columns of the matrix $\boldsymbol{W}$ and $h_{k,t}^z$ are the diagonal elements of the matrix $\boldsymbol{H}_t^z$. The difference between equations (8) and (10) is that the factors in (10) are uncorrelated but then, in the GO–GARCH model

it is not possible to have fewer factors than there are assets. This is possible in the O–GARCH model but at the cost of obtaining conditional covariance matrices with a reduced rank.

van der Weide (2002) constructs the linear mapping $\boldsymbol{W}$ by making use of the singular value decomposition of $E\boldsymbol{r}_t\boldsymbol{r}_t' = \boldsymbol{W}\boldsymbol{W}'$. That is,

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}\,,$$

where the columns of $\boldsymbol{U}$ hold the eigenvectors of $E\boldsymbol{r}_t\boldsymbol{r}_t'$ and the diagonal matrix $\boldsymbol{\Lambda}$ holds its eigenvalues, thus exploiting unconditional information only. Estimation of the orthogonal matrix $\boldsymbol{V}$ requires use of conditional information; see van der Weide (2002) for details.

Vrontos et al. (2003) have suggested a related model. They state their Full Factor (FF–) GARCH model as above but restrict the mapping $\boldsymbol{W}$ to be an $N \times N$ invertible triangular parameter matrix with ones on the main diagonal. Furthermore, the parameters in $\boldsymbol{W}$ are estimated directly using conditional information only. Assuming $\boldsymbol{W}$ to be triangular simplifies matters but is restrictive because, depending on the order of the components in the vector $\boldsymbol{r}_t$, certain relationships between the factors and the returns are ruled out.

Lanne and Saikkonen (2007) put forth yet another modelling proposal. In their Generalized Orthogonal Factor (GOF–) GARCH model the mapping $\boldsymbol{W}$ is decomposed using the polar decomposition:

$$\boldsymbol{W} = \boldsymbol{C}\boldsymbol{V}\,,$$

where $\boldsymbol{C}$ is a symmetric positive definite $N \times N$ matrix and $\boldsymbol{V}$ an orthogonal $N \times N$ matrix. Since $E\boldsymbol{r}_t\boldsymbol{r}_t' = \boldsymbol{W}\boldsymbol{W}' = \boldsymbol{C}\boldsymbol{C}'$, the matrix $\boldsymbol{C}$ can be estimated making use of the spectral decomposition $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}'$, where the columns of $\boldsymbol{U}$ are the eigenvectors of $E\boldsymbol{r}_t\boldsymbol{r}_t'$ and the diagonal matrix $\boldsymbol{\Lambda}$ contains its eigenvalues, thus using unconditional information only. Estimation of $\boldsymbol{V}$ requires the use of conditional information, see Lanne and Saikkonen (2007) for details.

An important aspect of the GOF–GARCH model is that some of the factors can be conditionally homoskedastic. In addition to being parsimonious, this allows the model to include not only systematic but also idiosyncratic components of risk. Suppose K ($\leq N$) of the factors are heteroskedastic, while the remaining $N - K$ factors are homoskedastic. Without loss of generality we can assume that the K first elements of $\boldsymbol{z}_t$ are the heteroskedastic ones, in which case this restriction is imposed by setting that the $N - K$ last diagonal elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ in (9) equal to zero. This results in the conditional covariance matrix of $\boldsymbol{r}_t$ of the following form (ref. eq. (10)):

$$\boldsymbol{H}_t = \sum_{k=1}^{K} \boldsymbol{w}_{(k)} \boldsymbol{w}'_{(k)} h^z_{k,t} + \sum_{k=K+1}^{N} \boldsymbol{w}_{(k)} \boldsymbol{w}'_{(k)}$$
$$= \sum_{k=1}^{K} \boldsymbol{w}_{(k)} \boldsymbol{w}'_{(k)} h^z_{k,t} + \boldsymbol{\Omega}. \tag{11}$$

The expression (11) is very similar to the one in (8), but there are two important differences. In (11) the factors are uncorrelated, whereas in (8), as already pointed out, this is not generally the case. The role of $\boldsymbol{\Omega}$ in (11) is also different from that of $\boldsymbol{\Omega}$ in (8). In the factor ARCH model $\boldsymbol{\Omega}$ is required to be a positive semi-definite matrix and it has no particular interpretation. For comparison, the matrix $\boldsymbol{\Omega}$ in the GOF–GARCH model has a reduced rank directly related to the number of heteroskedastic factors. Furthermore, it is closely related to the unconditional covariance matrix of $\boldsymbol{r}_t$. This results to the model being possibly considerably more parsimonious than the factor ARCH model; for details and a more elaborate discussion, see Lanne and Saikkonen (2007). Therefore, the GOF–GARCH model can be seen as combining the advantages of both the factor models (having a reduced number of heteroskedastic factors) and the orthogonal models (relative ease of estimation due to the orthogonality of factors).

2.3 Models of conditional variances and correlations

Correlation models are based on the decomposition of the conditional covariance matrix into conditional standard deviations and correlations. The simplest multivariate correlation model that is nested in the other conditional correlation models, is the Constant Conditional Correlation (CCC–) GARCH model of Bollerslev (1990). In this model, the conditional correlation matrix is time-invariant, so the conditional covariance matrix can be expressed as follows:

$$\boldsymbol{H}_t = \boldsymbol{D}_t \boldsymbol{P} \boldsymbol{D}_t \,, \tag{12}$$

where $\boldsymbol{D}_t = \mathrm{diag}(h_{1t}^{1/2}, \ldots, h_{Nt}^{1/2})$ and $\boldsymbol{P} = [\rho_{ij}]$ is positive definite with $\rho_{ii} = 1$, $i = 1, \ldots, N$. This means that the off-diagonal elements of the conditional covariance matrix are defined as follows:

$$[\boldsymbol{H}_t]_{ij} = h_{it}^{1/2} h_{jt}^{1/2} \rho_{ij}, \quad i \neq j \,,$$

where $1 \leq i, j \leq N$. The models for the processes $\{r_{it}\}$ are members of the class of univariate GARCH models. They are most often modelled as the GARCH(p,q) model, in which case the conditional variances can be written in a vector form

$$\boldsymbol{h}_t = \boldsymbol{\omega} + \sum_{j=1}^{q} \boldsymbol{A}_j \boldsymbol{r}_{t-j}^{(2)} + \sum_{j=1}^{p} \boldsymbol{B}_j \boldsymbol{h}_{t-j}\,, \tag{13}$$

where $\boldsymbol{\omega}$ is $N \times 1$ vector, $\boldsymbol{A_j}$ and $\boldsymbol{B_j}$ are diagonal $N \times N$ matrices, and $\boldsymbol{r}_t^{(2)} = \boldsymbol{r}_t \odot \boldsymbol{r}_t$. When the conditional correlation matrix $\boldsymbol{P}$ is positive definite and the elements of $\boldsymbol{\omega}$ and the diagonal elements of $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ positive, the conditional covariance matrix $\boldsymbol{H}_t$ is positive definite. Positivity of the diagonal elements of $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ is not, however, necessary for $\boldsymbol{P}$ to be positive definite unless $p = q = 1$, see Nelson and Cao (1992) for discussion of positivity conditions for h_{it} in univariate GARCH(p,q) models.

An extension to the CCC–GARCH model was introduced by Jeantheau (1998). In this Extended CCC– (ECCC–) GARCH model the assumption that the matrices $\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ in (13) are diagonal is relaxed. This allows the past squared returns and variances of all series to enter the individual conditional variance equations. For instance, in the first-order ECCC–GARCH model, the ith variance equation is

$$h_{it} = \omega_i + a_{11} r_{1,t-1}^2 + \ldots + a_{1N} r_{N,t-1}^2 + b_{11} h_{1,t-1} + \ldots + b_{1N} h_{N,t-1}, \quad i = 1, \ldots, N.$$

An advantage of this extension is that it allows a considerably richer autocorrelation structure for the squared observed returns than the standard CCC–GARCH model. For example, in the univariate GARCH(1,1) model the autocorrelations of the squared observations decrease exponentially from the first lag. In the first-order ECCC–GARCH model, the same autocorrelations need not have a monotonic decline from the first lag. This has been shown by He and Teräsvirta (2004) who considered the fourth-moment structure of first- and second-order ECCC–GARCH models.

The estimation of MGARCH models with constant correlations is computationally attractive. Because of the decomposition (12), the log-likelihood in (3) has the following simple form:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) = c - (1/2) \sum_{t=1}^{T} \sum_{i=1}^{N} \ln |h_{it}| - (1/2) \sum_{t=1}^{T} \log |\boldsymbol{P}| - (1/2) \sum_{t=1}^{T} \boldsymbol{r}_t' \boldsymbol{D}_t^{-1} \boldsymbol{P}^{-1} \boldsymbol{D}_t^{-1} \boldsymbol{r}_t. \tag{14}$$

From (14) it is apparent that during estimation, one has to invert the conditional correlation matrix only once per iteration. The number of parameters in the CCC– and ECCC–GARCH models, in addition to the ones in the univariate GARCH equations, equals $N(N-1)/2$ and covariance stationarity is ensured if the roots of $\det(\boldsymbol{I} - \sum_{j=1}^{q} \boldsymbol{A}_j \lambda^j - \sum_{j=1}^{p} \boldsymbol{B}_j \lambda^j) = 0$ lie outside the unit circle.

Although the CCC–GARCH model is in many respects an attractive parameterization, empirical studies have suggested that the assumption of constant conditional correlations may be too restrictive. The model may therefore be generalized by retaining the previous decomposition but making the conditional correlation matrix in (12) time-varying. Thus,

$$\boldsymbol{H}_t = \boldsymbol{D}_t \boldsymbol{P}_t \boldsymbol{D}_t. \tag{15}$$

In conditional correlation models defined through (15), positive definiteness of $\boldsymbol{H}_t$ follows if, in addition to the conditional variances h_{it}, $i = 1, \ldots, N$, being well-defined, the conditional correlation matrix $\boldsymbol{P}_t$ is positive definite at each point in time. Compared to the CCC–GARCH models, the advantage of numerically simple estimation is lost, as the correlation matrix has to be inverted for each t during every iteration.

Due to the intuitive interpretation of correlations, there exist a vast number of proposals for specifying $\boldsymbol{P}_t$. Tse and Tsui (2002) imposed GARCH type of dynamics on the conditional correlations. The conditional correlations in their Varying Correlation (VC–) GARCH model are functions of the conditional correlations of the previous period and a set of estimated correlations. More specifically,

$$\boldsymbol{P}_t = (1 - a - b)\boldsymbol{S} + a\boldsymbol{S}_{t-1} + b\boldsymbol{P}_{t-1}\,,$$

where $\boldsymbol{S}$ is a constant, positive definite parameter matrix with ones on the diagonal, a and b are non-negative scalar parameters such that $a + b \leq 1$, and $\boldsymbol{S}_{t-1}$ is a sample correlation matrix of the past M standardized residuals $\hat{\boldsymbol{\varepsilon}}_{t-1}, \ldots, \hat{\boldsymbol{\varepsilon}}_{t-M}$, where $\hat{\boldsymbol{\varepsilon}}_{t-j} = \widehat{\boldsymbol{D}}_{t-j}^{-1}\boldsymbol{r}_{t-j}$, $j = 1, \ldots, M$. The positive definiteness of $\boldsymbol{P}_t$ is ensured by construction if $\boldsymbol{P}_0$ and $\boldsymbol{S}_{t-1}$ are positive definite. A necessary condition for the latter to hold is $M \geq N$. The definition of the 'intercept' $1 - a - b$ corresponds to the idea of 'variance targeting' in Engle and Mezrich (1996).

Kwan et al. () proposed a threshold extension to the VC–GARCH model. Within each regime, indicated by the value of an indicator or threshold variable, the model has a VC–GARCH specification. Specifically, the authors partition the real line into R subintervals, $r_0 = -\infty < l_1 < \ldots < l_{R-1} < l_R = \infty$, and define an indicator variable $s_t \in \mathcal{F}_{t-1}^*$, the extended information set. The rth regime is defined by $l_{r-1} < s_t \leq l_r$, and both the univariate GARCH models and the dynamic correlations have regime-specific parameters. Kwan et al. () also apply the same idea to the BEKK model and discuss estimation of the number of regimes. In order to estimate the model consistently, one has to make sure that each regime contains a sufficient number of observations.

Engle (2002) introduced a Dynamic Conditional Correlation (DCC–) GARCH model whose dynamic conditional correlation structure is similar to that of the VC–GARCH model. Engle considered a dynamic matrix process

$$\boldsymbol{Q}_t = (1 - a - b)\boldsymbol{S} + a\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}' + b\boldsymbol{Q}_{t-1}\,,$$

where a is a positive and b a non-negative scalar parameter such that $a+b<1$, $\boldsymbol{S}$ is the unconditional correlation matrix of the standardized errors $\boldsymbol{\varepsilon}_t$, and $\boldsymbol{Q}_0$ is positive definite. This process ensures positive definiteness but does not generally produce valid correlation matrices. They are obtained by rescaling $\boldsymbol{Q}_t$ as follows:

$$\boldsymbol{P}_t = (\boldsymbol{I} \odot \boldsymbol{Q}_t)^{-1/2} \boldsymbol{Q}_t (\boldsymbol{I} \odot \boldsymbol{Q}_t)^{-1/2}.$$

Both the VC– and the DCC–GARCH model extend the CCC–GARCH model, but do it with few extra parameters. In each correlation equation, the number of parameters is $N(N-1)/2+2$ for the VC–GARCH model and two for in the DCC–GARCH one. This is a strength of these models but may also be seen as a weakness when N is large, because all $N(N-1)/2$ correlation processes are restricted to have the same dynamic structure.

To avoid this limitation, various generalizations of the DCC–GARCH model have been proposed. Billio and Carporin (2006) suggested a model that imposes a BEKK structure on the conditional correlations. Their Quadratic Flexible DCC (GFDCC) GARCH model where the matrix process Q_t is defined as

$$\boldsymbol{Q}_t = \boldsymbol{C}'\boldsymbol{S}\boldsymbol{C} + \boldsymbol{A}'\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}'_{t-1}\boldsymbol{A} + \boldsymbol{B}'\boldsymbol{Q}_{t-1}\boldsymbol{B}\,, \tag{16}$$

where the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are symmetric, and $\boldsymbol{S}$ is the unconditional covariance matrix of the standardized errors $\boldsymbol{\varepsilon}_t$. To obtain stationarity, $\boldsymbol{C}'\boldsymbol{S}\boldsymbol{C}$ has to be positive definite and the eigenvalues of $\boldsymbol{A}+\boldsymbol{B}$ must be less than one in modulus. The number of parameters governing the correlations in the GFDCC–GARCH model in its fully general form is $3N(N+1)/2$ which is unfeasible in large systems. The authors therefore suggested several special cases: One is to group the assets according to their properties, sector, or industry and restricting the coefficient matrices to be block diagonal following the partition. Another is to restrict the coefficient matrices to be diagonal with possibly suitable partition.

Cappiello et al. (2006) generalized the DCC–GARCH model in a similar manner, but also including asymmetric effects. In their Asymmetric Generalized DCC (AG–DCC) GARCH model the dynamics of $\boldsymbol{Q}_t$ is the following:

$$\begin{aligned}\boldsymbol{Q}_t = {} & (\boldsymbol{S} - \boldsymbol{A}'\boldsymbol{S}\boldsymbol{A} - \boldsymbol{B}'\boldsymbol{S}\boldsymbol{B} - \boldsymbol{G}'\boldsymbol{S}^{-}\boldsymbol{G}) + \boldsymbol{A}'\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}'_{t-1}\boldsymbol{A} \\ & + \boldsymbol{B}'\boldsymbol{Q}_{t-1}\boldsymbol{B} + \boldsymbol{G}'\boldsymbol{\varepsilon}^{-}_{t-1}\boldsymbol{\varepsilon}^{-\prime}_{t-1}\boldsymbol{G}\,,\end{aligned} \tag{17}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{G}$ are $N \times N$ parameter matrices, $\boldsymbol{\varepsilon}^{-} = \mathbb{I}_{\{\boldsymbol{\varepsilon}_t<0\}} \odot \boldsymbol{\varepsilon}_t$, where $\mathbb{I}$ is an indicator function, and $\boldsymbol{S}$ and $\boldsymbol{S}^{-}$ are the unconditional covariance matrices of $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}^{-}_t$, respectively. Again, the number of parameters increases rapidly with the dimension of the model, and restricted versions, such as diagonal, scalar, and symmetric, were suggested.

In the VC–GARCH as well as the DCC–GARCH model, the dynamic structure of the time-varying correlations is a function of past returns. There is another class of models that allows the dynamic structure of the correlations to be controlled by an exogenous variable. This variable may be either

an observable variable, a combination of observable variables, or a latent variable that represents factors that are difficult to quantify. One may argue that these models are not pure vector GARCH models because the conditioning set in them can be larger than in VC–GARCH or DCC–GARCH models. The first one of these models to be considered here is the Smooth Transition Conditional Correlation (STCC–) GARCH model.

In the STCC–GARCH model of Silvennoinen and Teräsvirta (2005), the conditional correlation matrix varies smoothly between two extreme states according to a transition variable. The following dynamic structure is imposed on the conditional correlations:

$$\boldsymbol{P}_t = (1 - G(s_t))\boldsymbol{P}_{(1)} + G(s_t)\boldsymbol{P}_{(2)}\,,$$

where $\boldsymbol{P}_{(1)}$ and $\boldsymbol{P}_{(2)}$, $\boldsymbol{P}_{(1)} \neq \boldsymbol{P}_{(2)}$, are positive definite correlation matrices that describe the two extreme states of correlations, and $G(\cdot) : \mathbb{R} \to (0,1)$, is a monotonic function of an observable transition variable $s_t \in \mathcal{F}^*_{t-1}$. The authors define $G(\cdot)$ as the logistic function

$$G(s_t) = \left(1 + e^{-\gamma(s_t - c)}\right)^{-1}, \quad \gamma > 0\,, \tag{18}$$

where the parameter γ determines the velocity and c the location of the transition. In addition to the univariate variance equations, the STCC–GARCH model has $N(N-1)+2$ parameters. The sequence $\{\boldsymbol{P}_t\}$ is a sequence of positive definite matrices because each $\boldsymbol{P}_t$ is a convex combination of two positive definite correlation matrices. The transition variable s_t is chosen by the modeller to suit the application at hand. If there is uncertainty about an appropriate choice of s_t, testing the CCC–GARCH model can be used as tool for judging the relevance of a given transition variable to the dynamic conditional correlations. A special case of the STCC–GARCH model is obtained when the transition variable is calendar time. The Time Varying Conditional Correlation (TVCC–) GARCH model was in its bivariate form introduced by Berben and Jansen (2005).

A recent extension of the STCC–GARCH model, the Double Smooth Transition Conditional Correlation (DSTCC–) GARCH model by Silvennoinen and Teräsvirta (2007) allows for another transition around the first one:

$$\begin{aligned}\boldsymbol{P}_t = {} & (1 - G_2(s_{2t}))\left\{(1 - G_1(s_{1t}))\boldsymbol{P}_{(11)} + G_1(s_{1t})\boldsymbol{P}_{(21)}\right\} \\ & + G_2(s_{2t})\left\{(1 - G_1(s_{1t}))\boldsymbol{P}_{(12)} + G_1(s_{1t})\boldsymbol{P}_{(22)}\right\}.\end{aligned} \tag{19}$$

For instance, one of the transition variables can simply be calendar time. If this is the case, one has the Time Varying Smooth Transition Conditional Correlation (TVSTCC–) GARCH model that nests the STCC–GARCH as well as the TVCC–GARCH model. The interpretation of the extreme states is the following: At the beginning of the sample, $\boldsymbol{P}_{(11)}$ and $\boldsymbol{P}_{(21)}$ are the two extreme states between which the correlations vary according to the transition

variable s_{1t} and similarly, $\boldsymbol{P}_{(12)}$ and $\boldsymbol{P}_{(22)}$ are the corresponding states at the end of the sample. The TVSTCC–GARCH model allows the extreme states, constant in the STCC–GARCH framework, to be time-varying, which introduces extra flexibility when modelling long time series. The number of parameters, excluding the univariate GARCH equations, is $2N(N-1)+4$ which restricts the use of the model in very large systems.

The Regime Switching Dynamic Correlation (RSDC–) GARCH model introduced by Pelletier (2006) falls somewhere between the models with constant correlations and the ones with correlations changing continuously at every period. The model imposes constancy of correlations within a regime while the dynamics enter through switching regimes. Specifically,

$$\boldsymbol{P}_t = \sum_{r=1}^{R} \mathbb{I}_{\{\Delta_t = r\}} \boldsymbol{P}_{(r)} \,,$$

where Δ_t is a (usually first-order) Markov chain independent of $\boldsymbol{\eta}_t$ that can take R possible values and is governed by a transition probability matrix $\boldsymbol{\Pi}$, $\mathbb{I}$ is the indicator function, and $\boldsymbol{P}_{(r)}$, $r = 1, \ldots, R$, are positive definite regime-specific correlation matrices. The correlation component of the model has $RN(N-1)/2 - R(R-1)$ parameters. A version that involves fewer parameters is obtained by restricting the R possible states of correlations to be linear combinations of a state of zero correlations and that of possibly high correlations. Thus,

$$\boldsymbol{P}_t = (1 - \lambda(\Delta_t))\boldsymbol{I} + \lambda(\Delta_t)\boldsymbol{P} \,,$$

where $\boldsymbol{I}$ is the identity matrix ('no correlations'), $\boldsymbol{P}$ is a correlation matrix representing the state of possibly high correlations, and $\lambda(\cdot) : \{1, \ldots, R\} \to [0, 1]$ is a monotonic function of Δ_t. The number of regimes R is not a parameter to be estimated. The conditional correlation matrices are positive definite at each point in time by construction both in the unrestricted and restricted version of the model. If N is not very small, Pelletier (2006) recommends two-step estimation. First estimate the parameters of the GARCH equations and, second, conditionally on these estimates, estimate the correlations and the switching probabilities using the EM algorithm of Dempster et al. (1977).

2.4 Nonparametric and semiparametric approaches

Non- and semiparametric models form an alternative to parametric estimation of the conditional covariance structure. These approaches have the advantage of not imposing a particular (possibly misspecified) structure on the data. One advantage of at least a few fully parametric multivariate GARCH

models is, however, that they offer an interpretation of the dynamic structure of the conditional covariance or correlation matrices. Another is that the quasi-maximum likelihood estimator is consistent when the errors are assumed multivariate normal. However, there may be considerable efficiency losses in finite samples if the returns are not normally distributed. Semiparametric models combine the advantages of a parametric model in that they reach consistency and sustain the interpretability, and those of a nonparametric model which is robust against distributional misspecification. Nonparametric models, however, suffer from the 'curse of dimensionality': due to the lack of data in all directions of the multidimensional space, the performance of the local smoothing estimator deteriorates quickly as the dimension of the conditioning variable increases, see Stone (1980). For this reason, it has been of interest to study methods for dimension-reduction or to use a single, one-dimensional conditioning variable. Developments in semi- and nonparametric modelling are discussed in detail in two separate chapters of this Handbook, see Linton (2008) andFranke et al. (2008).

One alternative is to specify a parametric model for the conditional covariance structure but estimate the error distribution nonparametrically, thereby attempting to offset the efficiency loss of the quasi-maximum likelihood estimator compared to the maximum likelihood estimator of the correctly specified model. In the semiparametric model of Hafner and Rombouts (2007) the data are generated by any particular parametric MGARCH model and the error distribution is unspecified but estimated nonparametrically. Their approach leads to the log-likelihood

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}) = c - (1/2) \sum_{t=1}^{T} \ln |\boldsymbol{H}_t| + \sum_{t=1}^{T} \ln g(\boldsymbol{H}_t^{-1/2} \boldsymbol{r}_t)\,, \tag{20}$$

where $g(\cdot)$ is an unspecified density function of the standardized residuals $\boldsymbol{\eta}_t$ such that $E[\boldsymbol{\eta}_t] = \mathbf{0}$ and $E[\boldsymbol{\eta}_t \boldsymbol{\eta}_t'] = \boldsymbol{I}$. This model may be seen as a multivariate extension of the semiparametric GARCH model by Engle and Gonzalez-Rivera (1991). A flexible error distribution blurs the line between the parametric structure and the distribution of the errors. For example, if the correlation structure of a semiparametric GARCH model is misspecified, a nonparametric error distribution may absorb some of the misspecification. The nonparametric method for estimating the density g is discussed in detail in Hafner and Rombouts (2007). They assume that g belongs to the class of spherical distributions. Even with this restriction their semiparametric estimator remains more efficient than the maximum likelihood estimator if the errors z_t are non-normal.

Long and Ullah (2005) introduce an approach similar to the previous one in that the model is based on any parametric MGARCH model. After estimating a parametric model, the estimated standardized residuals $\widehat{\boldsymbol{\eta}}_t$ are extracted. When the model is not correctly specified, these residuals may have some structure in them, and Long and Ullah (2005) use nonparamet-

ric estimation to extract this information. This is done by estimating the conditional covariance matrix using the Nadaraya-Watson estimator

$$\boldsymbol{H}_t = \widehat{\boldsymbol{H}}_t^{1/2} \frac{\sum_{\tau=1}^{T} \widehat{\boldsymbol{\eta}}_\tau \widehat{\boldsymbol{\eta}}_\tau' K_h(s_\tau - s_t)}{\sum_{\tau=1}^{T} K_h(s_\tau - s_t)} \widehat{\boldsymbol{H}}_t^{1/2}, \tag{21}$$

where $\widehat{\boldsymbol{H}}_t$ is the conditional covariance matrix estimated parametrically from an MGARCH model, $s_t \in \mathcal{F}_{t-1}^*$ is an observable variable that the model is conditioned on, $\hat{\boldsymbol{\varepsilon}}_t = \widehat{\boldsymbol{D}}_t^{-1} \boldsymbol{r}_t$, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is the bandwidth parameter. Positive definiteness of $\widehat{\boldsymbol{H}}_t$ ensures positive definiteness of the semiparametric estimator $\boldsymbol{H}_t$.

In the Semi-Parametric Conditional Correlation (SPCC–) GARCH model of Hafner et al. (2005), the conditional variances are modelled parametrically by any choice of univariate GARCH model, where $\hat{\boldsymbol{\varepsilon}}_t = \widehat{\boldsymbol{D}}_t^{-1} \boldsymbol{r}_t$ is the vector consisting of the standardized residuals. The conditional correlations $\boldsymbol{P}_t$ are then estimated using a transformed Nadaraya-Watson estimator:

$$\boldsymbol{P}_t = (\boldsymbol{I} \odot \boldsymbol{Q}_t)^{-1/2} \boldsymbol{Q}_t (\boldsymbol{I} \odot \boldsymbol{Q}_t)^{-1/2},$$

where

$$\boldsymbol{Q}_t = \frac{\sum_{\tau=1}^{T} \hat{\boldsymbol{\varepsilon}}_\tau \hat{\boldsymbol{\varepsilon}}_\tau' K_h(s_\tau - s_t)}{\sum_{\tau=1}^{T} K_h(s_\tau - s_t)}. \tag{22}$$

In (22), $s_t \in \mathcal{F}_{t-1}^*$ is a conditioning variable, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is the bandwidth parameter.

Long and Ullah (2005) also suggest estimating the covariance structure in a fully nonparametric fashion so that the model is not an MGARCH model, but merely a parameter-free multivariate volatility model. The estimator of the conditional covariance matrix is

$$\boldsymbol{H}_t = \frac{\sum_{\tau=1}^{T} \boldsymbol{r}_\tau \boldsymbol{r}_\tau' K_h(s_\tau - s_t)}{\sum_{\tau=1}^{T} K_h(s_\tau - s_t)},$$

where $s_t \in \mathcal{F}_{t-1}^*$ is a conditioning variable, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is the bandwidth parameter. This approach ensures positive definiteness of $\boldsymbol{H}_t$.

The choice of the kernel function is not important and it could be any probability density function, whereas the choice of the bandwidth parameter h is crucial, see for instance Pagan and Ullah (1999), Sections 2.4.2 and 2.7. Long and Ullah (2005) consider the choice of an optimal fixed bandwidth, whereas Hafner et al. (2005) discuss a way of choosing a dynamic bandwidth parameter such that the bandwidth is larger in the tails of the marginal distribution of the conditioning variable s_t than it is in the mid-region of the distribution.

3 Statistical Properties

Statistical properties of multivariate GARCH models are only partially known. For the development of statistical estimation and testing theory, it would be desirable to have conditions for strict stationarity and ergodicity of a multivariate GARCH process, as well as conditions for consistency and asymptotic normality of the quasi-maximum likelihood estimator. The results that are available establish these properties in special cases and sometimes under strong conditions.

Jeantheau (1998) considers the statistical properties and estimation theory of the ECCC–GARCH model he proposes. He provides sufficient conditions for the existence of a weakly stationary and ergodic solution, which is also strictly stationary. This is done by assuming $E\boldsymbol{r}_t\boldsymbol{r}_t' < \infty$. It would be useful to have both a necessary and a sufficient condition for the existence of a strictly stationary solution, but this question remains open. Jeantheau (1998) also proves the strong consistency of the QML estimator for the ECCC–GARCH model. Ling and McAleer (2003) complement Jeantheau's results and also prove the asymptotic normality of the QMLE in the case of the ECCC–GARCH model. For the global asymptotic normality result, the existence of the sixth moment of $\boldsymbol{r}_t$ is required. The statistical properties of the second-order model are also investigated in He and Teräsvirta (2004), who provide sufficient conditions for the existence of fourth moments, and, furthermore, give expressions for the fourth moment as well as the autocorrelation function of squared observations as functions of the parameters.

Comte and Lieberman (2003) study the statistical properties of the BEKK model. Relying on a result in Boussama (1998), they give sufficient, but not necessary conditions for strict stationarity and ergodicity. Applying Jeantheau's results, they provide conditions for the strong consistency of the QMLE. Furthermore, they also prove the asymptotic normality of the QMLE, for which they assume the existence of the eighth moment of $\boldsymbol{r}_t$. The fourth-moment structure of the BEKK and VEC models is investigated by Hafner (2003), who gives necessary and sufficient conditions for the existence of the fourth moments and provides expressions for them. These expressions are not functions of the parameters of the model. As the factor models listed in Section 2.2 are special cases of the BEKK model, the results of Comte and Lieberman (2003) and Hafner (2003) also apply to them.

4 Hypothesis Testing in Multivariate GARCH Models

Testing the adequacy of estimated models is an important part of model building. Existing tests of multivariate GARCH models may be divided into two broad categories: general misspecification tests and specification tests. The purpose of the tests belonging to the former category is to check the

adequacy of an estimated model. Specification tests are different in the sense that they are designed to test the model against a parametric extension. Such tests have been constructed for the CCC–GARCH model, but obviously not for other models. We first review general misspecification tests.

4.1 General misspecification tests

Ling and Li (1997) derived a rather general misspecification test for multivariate GARCH models. It is applicable for many families of GARCH models. The test statistic has the following form:

$$Q(k) = T\boldsymbol{\gamma}_k' \widehat{\boldsymbol{\Omega}}_k^{-1} \boldsymbol{\gamma}_k \,, \tag{23}$$

where $\boldsymbol{\gamma}_k = (\gamma_1, ..., \gamma_k)'$ with

$$\gamma_j = \frac{\sum_{t=j+1}^{T} (\boldsymbol{r}_t' \widehat{\boldsymbol{H}}_t^{-1} \boldsymbol{r}_t - N)(\boldsymbol{r}_{t-j}' \widehat{\boldsymbol{H}}_{t-j}^{-1} \boldsymbol{r}_{t-j} - N)}{\sum_{t=1}^{T} (\boldsymbol{r}_t' \widehat{\boldsymbol{H}}_t^{-1} \boldsymbol{r}_t - N)^2} \tag{24}$$

$j = 1, \ldots, k$, $\widehat{\boldsymbol{H}}_t$ is an estimator of $\boldsymbol{H}_t$, and $\widehat{\boldsymbol{\Omega}}_k$ is the estimated covariance matrix of $\boldsymbol{\gamma}_k$, see Ling and Li (1997) for details. Under the null hypothesis H_0 that the GARCH model is correctly specified, that is, $\boldsymbol{\eta}_t \sim \text{IID}(\boldsymbol{0}, \boldsymbol{I})$, statistic (23) has an asymptotic χ^2 distribution with k degrees of freedom. Under H_0, $E\boldsymbol{r}_t' \boldsymbol{H}_t^{-1} \boldsymbol{r}_t = N$, and therefore the expression (24) is the jth-order sample autocorrelation between $\mathbf{r}_t' \boldsymbol{H}_t^{-1} \boldsymbol{r}_t = \boldsymbol{\eta}_t' \boldsymbol{\eta}_t$ and $\boldsymbol{r}_{t-j}' \boldsymbol{H}_{t-j}^{-1} \boldsymbol{r}_{t-j} = \boldsymbol{\eta}_{t-j}' \boldsymbol{\eta}_{t-j}$. The test may thus be viewed as a generalization of the portmanteau test of Li and Mak (1994) for testing the adequacy of a univariate GARCH model. In fact, when $N = 1$, (23) collapses into the Li and Mak statistic. The McLeod and Li (1983) statistic (Ljung-Box statistic applied to squared residuals), frequently used for evaluating GARCH models, is valid neither in the univariate nor in the multivariate case, see Li and Mak (1994) for the univariate case.

A simulation study by Tse and Tsui (1999) indicates that the Ling and Li portmanteau statistic (24) often has low power. The authors show examples of situations in which a portmanteau test based on autocorrelations of pairs of individual standardized residuals performs better. The drawback of this statistic is, however, that its asymptotic null distribution is unknown, and the statistic tends to be undersized. Each test is based only on a single pair of residuals.

Duchesne (2004) introduced the test which is a direct generalization of the portmanteau test of Li and Mak (1994) to the VEC–GARCH model (2). Let $\hat{\boldsymbol{\eta}}_t$ denote the maximum likelihood estimator of the error vector $\boldsymbol{\eta}_t$ in the VEC–GARCH model. The idea is to derive the asymptotic distribution of $\hat{\boldsymbol{c}}_j = \text{vech}(\hat{\boldsymbol{\eta}}_t \hat{\boldsymbol{\eta}}_{t-j}')$, where $j = 1, \ldots, k$, under the null hypothesis that

$\{\boldsymbol{\eta}_t\} \sim \text{NID}(\mathbf{0}, \boldsymbol{I})$. Once this has been done, one can combine the results and obtain the asymptotic null distribution of $\hat{\boldsymbol{c}}_{(k)} = (\hat{\boldsymbol{c}}_1', ..., \hat{\boldsymbol{c}}_k')'$, where the vectors $\hat{\boldsymbol{c}}_j$, $j = 1, \ldots, k$, are asymptotically uncorrelated when the null holds. This distribution is normal since the asymptotic distribution of each $\hat{\boldsymbol{c}}_k$ is normal. It follows that under the null hypothesis,

$$Q_D(k) = T\hat{\boldsymbol{c}}_{(k)}'\widehat{\boldsymbol{\Omega}}_k^{-1}\hat{\boldsymbol{c}}_{(k)} \xrightarrow{d} \chi^2(kN(N+1)/2)\,, \tag{25}$$

where $\widehat{\boldsymbol{\Omega}}_k$ is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{c}}_{(k)}$, defined in Duchesne (2004). This portmanteau test statistic collapses into the statistic of Li and Mak (1994) when $N = 1$. When $\{\boldsymbol{\eta}_t\} = \{\boldsymbol{\varepsilon}_t\}$, that is, when $\boldsymbol{H}_t \equiv \sigma^2\boldsymbol{I}$, the test (25) is a test of no multivariate ARCH. For $N = 1$, it is then identical to the well known portmanteau test of McLeod and Li (1983).

Yet another generalization of univariate tests can be found in Kroner and Ng (1998). Their misspecification tests are suitable for any multivariate GARCH model. Let

$$\boldsymbol{G}_t = \boldsymbol{r}_t\boldsymbol{r}_t' - \widehat{\boldsymbol{H}}_t\,,$$

where $\widehat{\boldsymbol{H}}_t$ has been estimated from a GARCH model. The elements of $\boldsymbol{G}_t = [g_{ijt}]$ are 'generalized residuals'. When the model is correctly specified, they form a matrix of martingale difference sequences with respect to the information set $\mathcal{F}_{t-1}$ that contains the past information until $t-1$. Thus any variable $x_s \in \mathcal{F}_{t-1}$ is uncorrelated with the elements of $\boldsymbol{G}_t$. Tests based on these misspecification indicators may then be constructed. This is done for each g_{ijt} separately. The suggested tests are generalizations of the sign-bias and size-bias tests of Engle and Ng (1993). The test statistics have an asymptotic χ^2 distribution with one degree of freedom when the null hypothesis is valid. If the dimension of the model is large and there are several misspecification indicators, the number of available tests may be very large.

Testing the adequacy of the CCC–GARCH model has been an object of interest since it was found that the assumption of constant correlations may sometimes be too restrictive in practice. Tse (2000) constructed a Lagrange multiplier (LM) test of the CCC–GARCH model against the following alternative, $\boldsymbol{P}_t$, to constant correlations:

$$\boldsymbol{P}_t = \boldsymbol{P} + \boldsymbol{\Delta} \odot \boldsymbol{r}_{t-1}\boldsymbol{r}_{t-1}'\,, \tag{26}$$

where $\boldsymbol{\Delta}$ is a symmetric parameter matrix with the main diagonal elements equal to zero. This means that the correlations are changing as functions of the previous observations. The null hypothesis is $H_0 : \boldsymbol{\Delta} = \mathbf{0}$ or, expressed as a vector equation, $\text{vecl}(\boldsymbol{\Delta}) = 0$.[2] Equation (26) does not define a particular alternative to conditional correlations as $\boldsymbol{P}_t$ is not necessarily a positive

[2] The operator vecl(·) stacks the columns of the strictly lower triangular part (excluding main diagonal elements) of its argument matrix.

definite matrix for every t. For this reason we interpret the test as a general misspecification test.

Bera and Kim (2002) present a test of a bivariate CCC–GARCH model against the alternative that the correlation coefficient is stochastic. The test is an Information Matrix test and as such an LM or score test. It is designed for a bivariate model, which restricts its usefulness in applications.

4.2 Tests for extensions of the CCC–GARCH model

The most popular extension of the CCC–GARCH model to-date is the DCC–GARCH model of Engle (2002). However, there does not seem to be any published work on developing tests of constancy of correlations directly against this model.

As discussed in Section 2.3, Silvennoinen and Teräsvirta (2005) extend the CCC–GARCH into a STCC–GARCH model in which the correlations fluctuate according to a transition variable. They construct an LM test for testing the constant correlation hypothesis against the smoothly changing correlations. Since the STCC–GARCH model is only identified when the correlations are changing, standard asymptotic theory is not valid. A good discussion of this problem can be found in Hansen (1996). The authors apply the technique in Luukkonen et al. (1988) in order to circumvent the identification problem. The null hypothesis is $\gamma = 0$ in (18), and a linearization of the correlation matrix $\boldsymbol{P}_t$ by the first-order Taylor expansion of (18) yields

$$\boldsymbol{P}_t^* = \boldsymbol{P}_{(1)} - s_t \boldsymbol{P}_{(2)}^*.$$

Under H_0, $\boldsymbol{P}_{(2)}^* = \boldsymbol{0}$ and the correlations are thus constant. The authors use this fact to build their LM-type test on the transformed null hypothesis H_0' : $\mathrm{vecl}(\boldsymbol{P}_{(2)}^*) = \boldsymbol{0}$ (the diagonal elements of $\boldsymbol{P}_{(2)}^*$ equal zero by definition). When H_0' holds, the test statistic has an asymptotic χ^2 distribution with $N(N-1)/2$ degrees of freedom. The authors also derive tests for the constancy hypothesis under the assumption that some of the correlations remain constant also under the alternative. Silvennoinen and Teräsvirta (2007) extend the Taylor expansion based test to the situation where the STCC–GARCH model is the null model and the alternative is the DSTCC–GARCH model. This test collapses into the test of the CCC–GARCH model against STCC–GARCH model when $G_1(s_{1t}) \equiv 1/2$ in (19).

5 An Application

In this section we compare some of the multivariate GARCH models considered in previous sections by fitting them to the same data set. In order to keep the comparison transparent, we only consider bivariate models. Our observations are the daily returns of S&P 500 index futures and 10-year bond futures from January 1990 to August 2003. This data set has been analyzed by Engle and Colacito (2006).[3] There is no consensus in the literature about how stock and long term bond returns are related. Historically, the long-run correlations have been assumed constant, an assumption that has led to contradicting conclusions because evidence for both positive and negative correlation has been found over the years (short-run correlations have been found to be affected, among other things, by news announcements). From a theoretical point of view, the long-run correlation between the two should be state-dependent, driven by macroeconomic factors such as growth, inflation, and interest rates. The way the correlations respond to these factors may, however, change over time.

For this reason it is interesting to see what the correlations between the two asset returns obtained from the models are and how they fluctuate over time. The focus of reporting results will therefore be on conditional correlations implied by the estimated models, that is, the BEKK–, GOF–, DCC–, DSTCC–, and SPCC–GARCH ones. In the last three models, the individual GARCH equations are simply symmetric first-order ones. The BEKK model is also of order one with $K = 1$. All computations have been performed using Ox, version 4.02, see Doornik (2002), and our own source code.

Estimation of the BEKK model turned out to be cumbersome. Convergence problems were encountered in numerical algorithms, but the iterations seemed to suggest diagonality of the coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. A diagonal BEKK model was eventually estimated without difficulty.

In the estimation of the GOF–GARCH model it is essential to obtain good initial estimates of the parameters; for details, see Lanne and Saikkonen (2007). Having done that, we experienced no difficulties in the estimation of this model with a single factor. Similarly, no convergence problems were encountered in the estimation of the DCC model of Engle (2002).

The DSTCC–GARCH model makes use of two transition variables. Because the DSTCC framework allows one to test for relevance of a variable, or variables, to the description of the dynamic structure of the correlations, we relied on the tests in Silvennoinen and Teräsvirta (2005) and Silvennoinen and Teräsvirta (2007) described in Section 4.2, to select relevant transition variables. Out of a multitude of variables, including both exogenous ones and variables constructed from the past observations, prices or returns, the

[3] The data set in Engle and Colacito (2006) begins in August 1988, but our sample starts from January 1990 because we also use the time series for a volatility index that is available only from that date onwards.

Chicago Board Options Exchange volatility index (VIX) that represents the market expectations of 30-day volatility turned out to lead to the strongest rejection of the null hypothesis, measured by the p-value. Calendar time seemed to be another obvious transition variable. As a result, the first-order TVSTCC–GARCH model was fitted to the bivariate data.

The semiparametric model of Hafner et al. (2005) also requires a choice of an indicator variable. Because the previous test results indicated that VIX is informative about the dynamics of the correlations, we chose VIX as the indicator variable. The SPCC–GARCH model was estimated using a standard kernel smoother with an optimal fixed bandwidth, see Pagan and Ullah (1999), Sections 2.4.2 and 2.7, for discussion on the choice of constant bandwidth.

The estimated conditional correlations are presented in Figure 1, whereas Table 1 shows the sample correlation matrix of the estimated time-varying correlations. The correlations from the diagonal BEKK model and the DCC–GARCH model are very strongly positively correlated, which is also obvious from Figure 1. The second-highest correlation of correlations is the one between the SPCC–GARCH and the GOF–GARCH model. The time-varying correlations are mostly positive during the 1990's and negative after the turn of the century. In most models, correlations seem to fluctuate quite randomly, but the TVSTCC–GARCH model constitutes an exception. This is due to the fact that one of the transition variables is calendar time. Interestingly, in the beginning of the period the correlation between the S&P 500 and bond futures is only mildly affected by the expected volatility (VIX) and remains positive. Towards the end, not only does the correlation gradually turn negative, but expected volatility seems to affect it very strongly. Rapid fluctuations are a consequence of the fact that the transition function with VIX as the transition variable has quite a steep slope. After the turn of the century, high values of VIX generate strongly negative correlations.

Although the estimated models do not display fully identical correlations, the general message in them remains more or less the same. It is up to the user to select the model he wants to use in portfolio management and forecasting. A way of comparing the models consists of inserting the estimated covariance matrices $\boldsymbol{H}_t$, $t = 1, \ldots, T$, into the Gaussian log-likelihood function (3) and calculate the maximum value of log-likelihood. These values for the estimated models appear in Table 1.

The models that are relatively easy to estimate seem to fit the data less well than the other models. The ones with a more complicated structure and, consequently, an estimation procedure that requires care, seem to attain higher likelihood values. However, the models do not make use of the same information set and, besides, they do not contain the same number of parameters. Taking this into account suggests the use of model selection criteria for assessing the performance of the models. Nevertheless, rankings by Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are the same as the likelihood-based ranking; see Table 1. Note that

in theory, rankings based on a model selection criterion favour the SPCC model. This is because no penalty is imposed on the nonparametric correlation estimates that improve the fit compared to constant correlations.

Nonnested testing as a means of comparison is hardly a realistic option here since the computational effort would be quite substantial. Out-of-sample forecasting would be another way of comparing models. However, the models involved would be multivariate and the quantities to be forecast would be measures of (unobserved) volatilities and cross-volatilities. This would give rise to a number of problems, beginning from defining the quantities to be forecast and appropriate loss functions, and from comparing forecast vectors instead of scalar forecasts. It appears that plenty of work remains to be done in that area.

	diag BEKK	GOF	DCC	TVSTCC	SPCC
diag BEKK	1.0000				
GOF	0.7713	1.0000			
DCC	0.9875	0.7295	1.0000		
TVSTCC	0.7577	0.7381	0.7690	1.0000	
SPCC	0.6010	0.8318	0.5811	0.7374	1.0000
log-likelihood	-6130	-6091	-6166	-6006	-6054
AIC	12275	12198	12347	12041	12120
BIC	12286	12211	12359	12062	12130

Table 1 Sample correlations of the estimated conditional correlations. The lower part of the table shows the log-likelihood values and the values of the corresponding model selection criteria.

6 Final Remarks

In this review, we have considered a number of multivariate GARCH models and highlighted their features. It is obvious that the original VEC model contains too many parameters to be easily applicable, and research has been concentrated on finding parsimonious alternatives to it. Two lines of development are visible. First, there are the attempts to impose restrictions on the parameters of the VEC model. The BEKK model and the factor models are examples of this. Second, there is the idea of modelling conditional covariances through conditional variances and correlations. It has led to a number of new models, and this family of conditional correlation models appears to be quite popular right now. The conditional correlation models are easier to estimate than many of their counterparts and their parameters (correlations) have a natural interpretation.

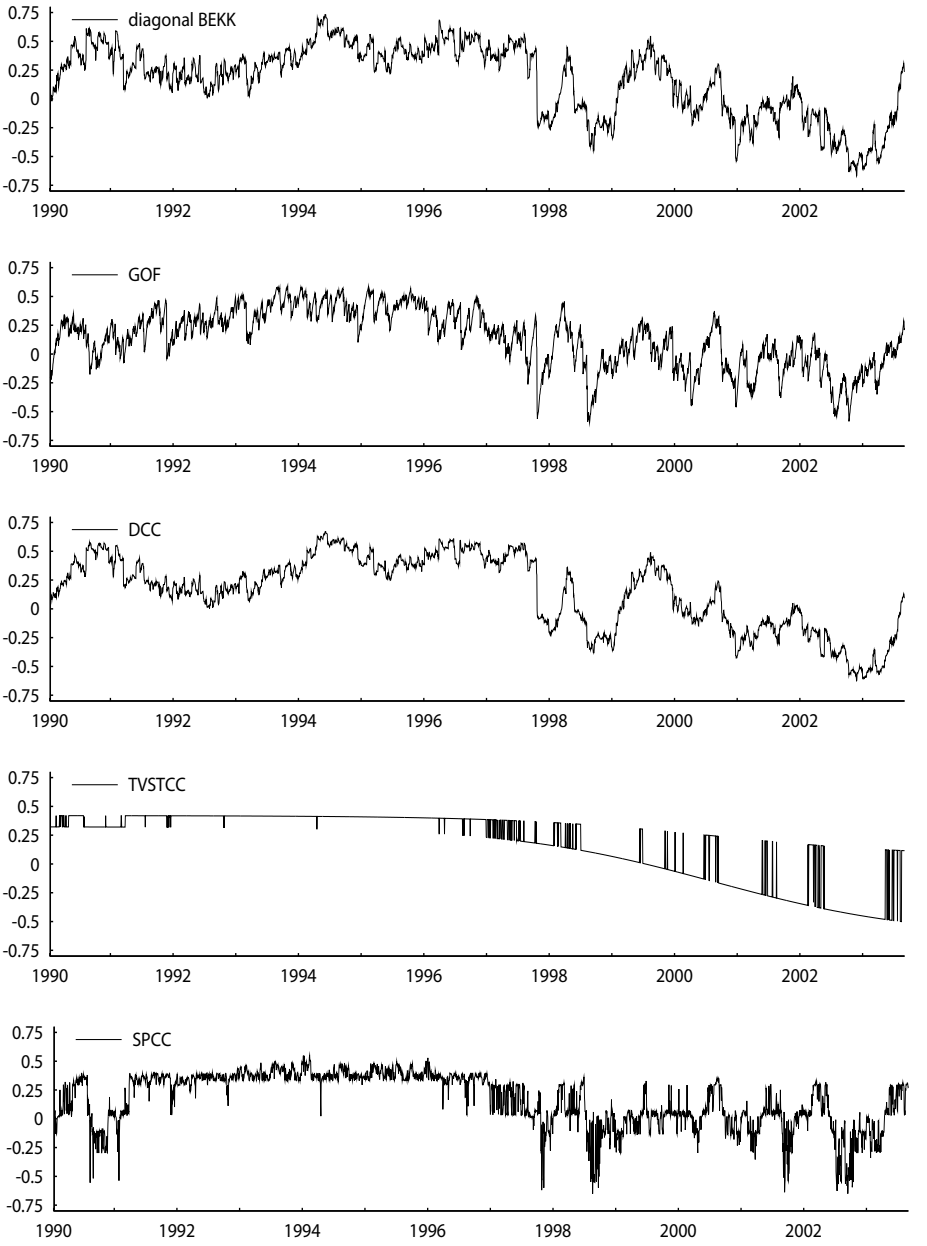

Fig. 1 Conditional correlations implied by the estimated models: Diagonal BEKK, GOF–GARCH, DCC–GARCH, TVSTCC–GARCH, and SPCC–GARCH.

As previously discussed, there is no statistical theory covering all MGARCH models. This may be expected, since models in the two main categories differ substantially from each other. Progress has been made in some special situations, and these cases have been considered in previous sections.

Estimation of multivariate GARCH models is not always easy. BEKK models appear more difficult to estimate than the CCC-GARCH model and its generalizations. While it has not been the objective of this review to cover algorithms for performing the necessary iterations, Brooks et al. (2003) compared four software packages for estimating MGARCH models. They used a single bivariate dataset and only fitted a first-order VEC-GARCH model to the data. A remarkable thing is that already the parameter estimates resulting from these packages are quite different, not to mention standard deviation estimates. The estimates give rather different ideas of the persistence of conditional volatility. These differences do not necessarily say very much about properties of the numerical algorithms used in the packages. It is more likely that they reflect the estimation difficulties. The log-likelihood function may contain a large number of local maxima, and different starting-values may thus lead to different outcomes. See Silvennoinen (2008) for more discussion. The practitioner who may wish to use these models in portfolio management should be aware of these problems.

Not much has been done as yet to construct tests for evaluating MGARCH models. A few tests do exist, and a number of them have been considered in this review.

It may be that VEC and BEKK models, with the possible exception of factor models, have already matured and there is not much that can be improved. The situation may be different for conditional correlation models. The focus has hitherto been on modelling the possibly time-varying correlations. Less emphasis has been put on the GARCH equations that typically have been GARCH(1,1) specifications. Designing diagnostic tools for testing and improving GARCH equations may be one of the challenges for the future.

References

Alexander, C. O. and Chibumba, A. M. (1997): Multivariate orthogonal factor GARCH. University of Sussex Discussion Papers in Mathematics.

Bae, K.-H., Karolyi, G. A. and Stulz, R. M. (2003): A new approach to measuring financial contagion. *The Review of Financial Studies* **16**, 717–763.

Bauwens, L., Laurent, S. and Rombouts, J. V. K. (2006): Multivariate GARCH Models: A Survey. *Journal of Applied Econometrics* **21**, 79–109.

Bera, A. K. and Kim, S. (2002): Testing constancy of correlation and other specifications of the BGARCH model with an application to international equity returns. *Journal of Empirical Finance* **9**, 171–195.

Berben, R.-P. and Jansen, W. J. (2005): Comovement in international equity markets: A sectoral view. *Journal of International Money and Finance* **24**, 832–857.

Billio, M. and Caporin, M. (2006): A generalized dynamic conditional correlation model for portfolio risk evaluation. Unpublished manuscript, Ca'Foscari University of Venice, Department of Economics.

Bollerslev, T. (1990): Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* **72**, 498–505.

Bollerslev, T., Engle, R. F. and Nelson, D. B. (1994): ARCH Models. *In: Engle, R.F. and McFadden, D.L. (Eds.): Handbook of Econometrics* **4**, 2959–3038. North-Holland, Amsterdam.

Bollerslev, T., Engle R. F. and Wooldridge, J. M. (1988): A capital asset pricing model with time-varying covariances. *The Journal of Political Economy* **96**, 116–131.

Boussama, F. (1998): *Ergodicité, mélange et estimation dans le modelès GARCH.* PhD Thesis, Université 7 Paris.

Brooks, C., Burke S. P. and Persand G. (2003): Multivariate GARCH models: software choice and estimation issues. *Journal of Applied Econometrics* **18**, 725–734.

Cappiello, L., Engle, R. F. and Sheppard, K. (2006): Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics* **4**, 537–572.

Chib, S., Omori, Y. and Asai, M. (2008): Multivariate stochastic volatility. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 365–400. Springer, New York.

Comte, F. and Lieberman, O. (2003): Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis* **84**, 61–84.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1–38.

Diebold, F. X. and Nerlove, M. (1989): The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *Journal of Applied Econometrics* **4**, 1–21.

Doornik, J. A. (2002): *Object-Oriented Matrix Programming Using Ox* 3rd edition. Timberlake Consultants Press.

Duchesne, P. (2004): On matricial measures of dependence in vector ARCH models with applications to diagnostic checking. *Statistics and Probability Letters* **68**, 149–160.

Engle, R. F. (1982): Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **50**, 987–1007.

Engle, R. F. (2002): Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* **20**, 339–350.

Engle, R. F. and Colacito, R. (2006): Testing and valuing dynamic correlations for asset allocation. *Journal of Business and Economic Statistics* **24**, 238–253.

Engle, R. F. and Gonzalez-Rivera, G. (1991): Semiparametric ARCH models. *Journal of Business and Economic Statistics* **9**, 345–359.

Engle, R. F., Granger, C. W. J. and Kraft, D. (1984): Combining competing forecasts of inflation using a bivariate ARCH model. *Journal of Economic Dynamics and Control* **8**, 151–165.

Engle, R. F. and Kroner, K. F. (1995): Multivariate simultaneous generalized ARCH. *Econometric Theory* **11**, 122–150.

Engle, R. F. and Mezrich, J. (1996): GARCH for groups. *Risk* **9**, 36–40.

Engle, R. F. and Ng, V. K. (1993): Measuring and Testing the Impact of News on Volatility. *Journal of Finance* **48**, 1749–1777.

Engle, R. F., Ng, V. K. and Rothschild, M. (1990): Asset pricing with a factor ARCH covariance structure: empirical estimates for treasury bills. *Journal of Econometrics* **45**, 213–238.

Franke, J., Kreiss, J.-P. and Mammen, E. (2008): Nonparametric modeling in financial time series. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 926–952. Springer, New York.

Gouriéroux, C. (1997): ARCH Models and Financial Applications. Springer, Berlin.
Hafner, C. M. (2003): Fourth moment structure of multivariate GARCH models. *Journal of Financial Econometrics* **1**, 26–54.
Hafner, C. M. and Rombouts, J. V. K. (2007): Semiparametric multivariate volatility models. *Econometric Theory* **23**, 251–280.
Hafner, C. M., van Dijk, D. and Franses, P. H. (2005): Semi-parametric modelling of correlation dynamics. *In: Fomby, T., Hill, C. and Terrell, D. (Eds.): Advances in Econometrics* **20/A**, 59–103. Elsevier, Amsterdam.
Hansen, B. E. (1996): Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64**, 413–430.
Hansson, B. and Hordahl, P. (1998): Testing the conditional CAPM using multivariate GARCH–M. *Applied Financial Economics* **8**, 377–388.
He, C. and Teräsvirta, T. (2004): An extended constant conditional correlation GARCH model and its fourth-moment structure. *Econometric Theory* **20**, 904–926.
Jeantheau, T. (1998): Strong consistency of estimators for multivariate ARCH models. *Econometric Theory* **14**, 70–86.
Kawakatsu, H. (2006): Matrix exponential GARCH. *Journal of Econometrics* **134**, 95–128.
Kroner, K. F. and Ng, V. K. (1998): Modeling asymmetric comovements of asset returns. *The Review of Financial Studies* **11**, 817–844.
Kwan, C. K., Li, W. K. and Ng, K. (inpress): A multivariate threshold GARCH model with time-varying correlations. *Econometric Reviews* to appear.
Lanne, M. and Saikkonen, P. (2007): A multivariate generalized orthogonal factor GARCH model. *Journal of Business and Economic Statistics* **25**, 61–75.
Li, W. K. and Mak, T. K. (1994): On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *Journal of Time Series Analysis* **15**, 627–636.
Ling, S. and Li, W. K. (1997): Diagnostic checking of nonlinear multivariate time series with multivariate ARCH errors. *Journal of Time Series Analysis* **18**, 447–464.
Ling, S. and McAleer, M. (2003): Asymptotic theory for a vector ARMA–GARCH model. *Econometric Theory* **19**, 280–310.
Linton, O. B. (2008): Semiparametric and nonparametric ARCH modelling. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 156–167. Springer, New York.
Long, X. and Ullah, A. (2005): Nonparametric and semiparametric multivariate GARCH model. Unpublished manuscript.
Luukkonen, R., Saikkonen, P. and Teräsvirta, T. (1988): Testing linearity against smooth transition autoregressive models. *Biometrika* **75**, 491–499.
McLeod, A. I. and Li, W. K. (1983): Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**, 269–273.
Nelson, D. B. (1991): Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* **59**, 347–370.
Nelson, D. B. and Cao, C. Q. (1992): Inequality Constraints in the Univariate GARCH Model. *Journal of Business and Economic Statistics* **10**, 229–235.
Ng, L. (1991): Tests of the CAPM with time-varying covariances: a multivariate GARCH approach. *The Journal of Finance* **46**, 1507–1521.
Pagan, A. and Ullah, A. (1999): Nonparametric Econometrics. Cambridge University Press.
Palm, F. C. (1996): GARCH Models of Volatility. *In: Maddala, G. S. and Rao, C. R. (Eds.): Handbook of Statistics: Statistical Methods in Finance* **14**, 209–240. Elsevier, Amsterdam.
Pelletier, D. (2006): Regime switching for dynamic correlations. *Journal of Econometrics* **131**, 445–473.
Ross, S. A. (1976): The arbitrage theory of capital asset pricing. *Journal of Economic Theory* **13**, 341–360.
Sentana, E. (1998): The relation between conditionally heteroskedastic factor models and factor GARCH models. *Econometrics Journal* **1**, 1–9.

Shephard, N. G. (1996): Statistical Aspects of ARCH and Stochastic Volatility. *In: Cox, D. R., Hinkley, D. V. and Barndorff-Nielsen, O. E. (Eds.): Time Series Models in Econometrics, Finance and Other Fields*, 1–67. Chapman and Hall, London.

Silvennoinen, A. (2006): Numerical aspects of the estimation of multivariate GARCH models. Unpublished manuscript.

Silvennoinen, A. (2008): Numerical aspects of the estimation of multivariate GARCH models. QFRC Research Paper, University of Technology, Sydney.

Silvennoinen, A. and Teräsvirta, T. (2005): Multivariate autoregressive conditional heteroskedasticity with smooth transitions in conditional correlations. SSE/EFI Working Paper Series in Economics and Finance **577**.

Silvennoinen, A. and Teräsvirta, T. (2007): Modelling multivariate autoregressive conditional heteroskedasticity with the double smooth transition conditional correlation GARCH model. SSE/EFI Working Paper Series in Economics and Finance **625**.

Stone, C. (1980): Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* **8**, 1348–1360.

Tse, Y. K. (2000): A test for constant correlations in a multivariate GARCH model. *Journal of Econometrics* **98**, 107–127.

Tse, Y. K. and Tsui, K. C. (1999): A note on diagnosing multivariate conditional heteroscedasticity models. *Journal of Time Series Analysis*. **20**, 679–691.

Tse, Y. K. and Tsui, K. C. (2002): A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business and Economic Statistics* **20**, 351–362.

van der Weide, R. (2002): GO–GARCH: A multivariate generalized orthogonal GARCH model. *Journal of Applied Econometrics* **17**, 549–564.

Vrontos, I. D., Dellaportas, P. and Politis, D. N. (2003): A full-factor multivariate GARCH model. *Econometrics Journal* **6**, 312–334.

Stochastic Volatility: Origins and Overview

Neil Shephard and Torben G. Andersen *

Abstract Stochastic volatility is the main way time-varying volatility is modelled in financial markets. The development of stochastic volatility is reviewed, placing it in a modeling and historical context. Some recent trends in the literature are highlighted.

1 Introduction

Stochastic volatility (SV) models are used heavily within the fields of financial economics and mathematical finance to capture the impact of time-varying volatility on financial markets and decision making. The development of the subject has been highly multidisciplinary, with results drawn from financial economics, probability theory and econometrics blending to produce methods that aid our understanding of option pricing, efficient portfolio allocation and accurate risk assessment and management.

Time-varying volatility is endemic in financial markets. This has been known for a long time, with early comments including Mandelbrot (1963) and Fama (1965). It was also clear to the founding fathers of modern continuous-time finance that homogeneity was an unrealistic if convenient simplification, e.g. Black and Scholes (1972, p. 416), wrote "... there is evidence of non-stationarity in the variance. More work must be done to predict variances

Neil Shephard
Oxford-Man Institute, University of Oxford, Oxford, UK; Department of Economics, University of Oxford, Oxford, UK, e-mail: neil.shephard@economics.ox.ac.uk

Torben G. Andersen
Kellogg School of Management, Northwestern University, Evanston, IL; NBER, Cambridge, MA; and CREATES, Aarhus, Denmark, e-mail: t-andersen@kellogg.northwestern.edu

* The work of Andersen is supported by a grant from the NSF to the NBER and by CREATES funded by the Danish National Research Foundation.

DOI: 10.1007/978-3-540-71297-8_10,

using the information available." Heterogeneity has deep implications for the theory and practice of financial economics and econometrics. In particular, asset pricing theory implies that higher rewards are required as an asset is exposed to more systematic risk. Of course, such risks may change through time in complicated ways, and it is natural to build stochastic models for the temporal evolution in volatility and codependence across assets. This allows us to explain, for example, empirically observed departures from Black-Scholes-Merton option prices and understand why we should expect to see occasional dramatic moves in financial markets. More generally, they bring the application of financial economics closer to the empirical reality of the world we live in, allowing us to make better decisions, inspire new theory and improve model building.

Autoregressive conditionally heteroskedasticity (ARCH) processes are often described as SV, but we do not follow that nomenclature. The essential feature of ARCH models is that they explicitly model the conditional variance of returns given past returns observed by the econometrician. This one-step-ahead prediction approach to volatility modeling is very powerful, particularly in the field of risk management. It is convenient from an econometric viewpoint as it immediately delivers the likelihood function as the product of one-step-ahead predictive densities.

In the SV approach the predictive distribution of returns is specified indirectly, via the structure of the model, rather than directly. For a small number of SV models this predictive distribution can be calculated explicitly but, invariably, for empirically realistic representations it has to be computed numerically. This move away from direct one-step-ahead predictions has some advantages. In particular, in continuous time it is more convenient, and perhaps more natural, to model directly the volatility of asset prices as having its own stochastic process without worrying about the implied one-step-ahead distribution of returns recorded over an arbitrary time interval convenient for the econometrician, such as a day or a month. This does, however, raise some difficulties as the likelihood function for SV models is not directly available, much to the frustration of econometricians in the late 1980s and early 1990s.

Since the mid-1980s continuous-time SV has dominated the option pricing literature but early on econometricians struggled with the difficulties of estimating and testing these models. Only in the 1990s were novel simulation strategies developed to efficiently estimate SV models. These computationally intensive methods enable us, given enough coding and computing time, to efficiently estimate a broad range of fully parametric SV models. This has led to refinements of the models, with many earlier tractable models being rejected from an empirical viewpoint. The resulting enriched SV literature has brought us much closer to the empirical realities we face in financial markets.

From the late 1990s SV models have taken center stage in the econometric analysis of volatility forecasting using high-frequency data based on realized volatility and related concepts. The reason is that the econometric analysis of realized volatility is tied to continuous-time processes, so SV is

central. The close connection between SV and realized volatility has allowed financial econometricians to harness the enriched information set available through high-frequency data to improve, by an order of magnitude, the accuracy of their volatility forecasts over that traditionally offered by ARCH models based on daily observations. This has broadened the applications of SV into the important arena of risk assessment and asset allocation.

Below, we provide a selective overview of the SV literature. The exposition touches on models, inference, options pricing and realized volatility. The SV literature has grown organically, with a variety of contributions playing important roles for particular branches of the literature, reflecting the highly multidisciplinary nature of the research.

2 The Origin of SV Models

The modern treatment of SV is typically cast in continuous time, but many older contributions employ discrete-time models. Specifically, the early econometric studies tended to favor discrete-time specifications, while financial mathematicians and financial economists often cast the problems in a diffusive setting when addressing portfolio choice and derivatives pricing. In response, econometricians have more recently developed practical inference tools for continuous-time SV models. We start with a description of some important early studies cast in a discrete-time setting and then cover the continuous-time formulations.

A central intuition in the SV literature is that asset returns are well approximated by a mixture distribution where the mixture reflects the level of activity or news arrivals. Clark (1973) originates this approach by specifying asset prices as subordinated stochastic processes directed by the increments to an underlying activity variable. Ignoring mean returns and letting the directing process being independent of the return innovations he stipulates,

$$Y_i = X_{\tau_i}, \quad i = 0, 1, 2, ..., \tag{1}$$

where Y_i denotes the logarithmic asset price at time i and $y_i = Y_i - Y_{i-1}$ the corresponding continuously compounded return over $[i-1, i]$, X_i is a normally distributed random variable with mean zero, variance $\sigma_X^2 \cdot i$, and independent increments, and τ_i is a real-valued process initiated at $\tau_0 = 0$ with non-negative and non-decreasing sample paths, i.e., it constitutes a time change. Clark focuses on the case where the increments to τ_i represent independent draws from a stationary distribution with finite variance, implying the subordinated return process also has independent increments with zero mean. More generally, as long as the time change process is independent of the price innovations, the asset returns are serially uncorrelated, albeit dependent, even if the time change increments are not stationary or independent.

In fact, we have

$$y_i|(\tau_i - \tau_{i-1}) \sim N(0, \sigma_X^2 \cdot (\tau_i - \tau_{i-1})). \tag{2}$$

Thus, marginally, the asset returns follow a normal mixture, implying a symmetric but fat tailed distribution. The directing or mixing process, $\tau_t, t \geq 0$, is naturally interpreted as an indicator of the intensity of price-relevant information flow over the interval $[0, t]$. Specifications of this type are generally referred to as Mixture of Distributions Hypotheses (MDH). They induce heteroskedastic return volatility and, if the time-change process is positively serially correlated, also volatility clustering. Clark explores the i.i.d. time-change specification only and relates the time-change to trading volume. Many subsequent studies pursue the serially correlated volatility extension empirically and seek to identify observable market proxies for the latent time-change process. Complete specification of the joint dynamic distribution of return variation and related market variables allows for a more structurally oriented approach to stochastic volatility modeling, see, e.g., Epps and Epps (1976), Tauchen and Pitts (1983), Andersen (1996), and Leisenfeld (2001).

For future reference, it is convenient to cast the Clark formulation in equivalent continuous-time notation. To emphasize that the log-price process as specified is a martingale, we denote it M. We may then restate equation (1) in a manner which implies the identical distribution for discretely sampled data,

$$M_t = W_{\tau_t}, \quad t \geq 0, \tag{3}$$

where W is Brownian motion (BM) and W and τ are independent processes. Technically, as long as (for each t) $\mathrm{E}\sqrt{\tau_t} < \infty$, M is a martingale since this is necessary and sufficient to ensure that $\mathrm{E}\,|M_t| < \infty$.

Asset pricing theory asserts that securities exposed to systematic risk have expected positive excess returns relative to the risk-free interest rate. As a result, asset prices will not generally be martingales. Instead, assuming frictionless markets, a weak no-arbitrage condition implies that the asset price will be a special semimartingale, see, e.g., Back (1991). This leads to the more general formulation,

$$Y = Y_0 + A + M, \tag{4}$$

where the finite variation process, A, constitutes the expected mean return. If the asset represents a claim on the broader market portfolio, a simple and popular specification for A is $A_t = r_f t + \beta\tau_t$, with r_f denoting the risk-free rate and β representing a risk premium due to the undiversifiable variance risk. This means that the distributional MDH result in equation (2) generalizes to $Y_t|\tau_t \sim N(r_f t + \beta\tau_t, \tau_t)$.

Clark's main purpose was to advocate the MDH as an alternative to the empirically less attractive stable processes. Although his framework lends itself to the appropriate generalizations, he did not seek to accommodate the

persistence in return volatility. In fact, only about a decade later we find a published SV paper explicitly dealing with volatility clustering, namely Taylor (1982). Taylor models the risky part of returns as a product process,

$$m_i = M_i - M_{i-1} = \sigma_i \varepsilon_i. \tag{5}$$

ε is assumed to follow an autoregression with zero mean and unit variance, while σ is some non-negative process. He completes the model by assuming $\varepsilon \perp\!\!\!\perp \sigma$ and

$$\sigma_i = \exp(h_i/2), \tag{6}$$

where h is a non-zero mean Gaussian linear process. The leading example of this is the first order autoregression,

$$h_{i+1} = \mu + \phi\,(h_i - \mu) + \eta_i, \tag{7}$$

where η is a zero mean, Gaussian white noise process. In the modern SV literature the model for ε is typically simplified to an i.i.d. process, as the predictability of asset prices is incorporated in the A process rather than in M. The resulting model is now often called the log-normal SV model if ε is also assumed to be Gaussian. Finally, we note that M is a martingale as long as $\mathrm{E}\,(\sigma_i) < \infty$, which is satisfied for all models considered above if h is stationary.[2]

A key feature of SV, not discussed by Taylor, is that it can accommodate an asymmetric return-volatility relation, often termed a statistical leverage effect in reference to Black (1976), even if it is widely recognized that the asymmetry is largely unrelated to any underlying financial leverage. The effect can be incorporated in discrete-time SV models by negatively correlating the Gaussian ε_i and η_i so that the direction of returns impact future movements in the volatility process, with price drops associated with subsequent increases in volatility. Leverage effects also generate skewness, via the dynamics of the model, in the distribution of $(M_{i+s} - M_i)\,|\sigma_i$ for $s \geq 2$, although $(M_{i+1} - M_i)\,|\sigma_i$ continues to be symmetric. This is a major impetus for the use of these models in pricing of equity index options for which skewness appears endemic.

We now move towards a brief account of some early contributions to the continuous-time SV literature. In that context, it is useful to link the above exposition to the corresponding continuous-time specifications. The counter-

[2] Taylor's discussion of the product process was predated by a decade in the unpublished Rosenberg (1972). This remarkable paper appears to have been lost to the modern SV literature until recently, but is now available in Shephard (2005). Rosenberg introduces product processes, empirically demonstrating that time-varying volatility is partially predictable, and thus moving beyond Clark's analysis on this critical dimension. He also explores a variety of econometric methods for analyzing heteroskedasticity only reintroduced into the literature much later. Finally, he studies an SV model which in some respects is a close precursor of ARCH models even if he clearly does not recognize the practical significance of restrictions on his system that would lead to an ARCH representation.

part to the (cumulative) product process for the martingale component in equation (5) is given by the Itô stochastic integral representation,

$$M_t = \int_0^t \sigma_s \mathrm{d}W_s, \tag{8}$$

where the non-negative spot volatility σ is assumed to have càdlàg sample paths. Note that this allows for jumps in the volatility process. Moreover, SV models given by (8) have continuous sample paths even if σ does not. A necessary and sufficient condition for M to constitute a martingale is that $\mathrm{E}\sqrt{\int_0^t \sigma_s^2 \mathrm{d}s} < \infty$. The squared volatility process is often termed the spot variance. There is no necessity for σ and W to be independent, but when they are we obtain the important simplification that $M_t | \int_0^t \sigma_s^2 \mathrm{d}s \sim N\left(0, \int_0^t \sigma_s^2 \mathrm{d}s\right)$. This makes it evident that the structure is closely related to the MDH or time-change representation (3) of Clark. The directing process is labeled Integrated Variance, i.e., $IV_t = \int_0^t \sigma_s^2 \mathrm{d}s$, and arises naturally as a quantity of key interest in practical applications.

An early application of continuous-time SV models was the unpublished work by Johnson (1979) who studied option pricing using time-changing volatility. While this project evolved into Johnson and Shanno (1987), a well known paper on the use of continuous-time SV models for option pricing is Hull and White (1987) who allow the spot volatility process to follow a general diffusion. In their approach the spot variation process is given as the solution to a univariate stochastic differential equation,

$$\mathrm{d}\sigma^2 = \alpha(\sigma^2)\mathrm{d}t + \omega(\sigma^2)\mathrm{d}B, \tag{9}$$

where B is a second Brownian motion and $\alpha(.)$ and $\omega(.)$ are deterministic functions which can be specified quite generally but must ensure that σ^2 remains strictly positive. By potentially correlating the increments of W and B, Hull and White provide the first coherent leverage model in financial economics. They compute option prices by numerical means for the special case,

$$\mathrm{d}\sigma^2 = \alpha\sigma^2\mathrm{d}t + \omega\sigma^2\mathrm{d}B. \tag{10}$$

This formulation is quite similar to the so-called GARCH diffusion which arises as the diffusion limit of a sequence of GARCH(1,1) models, see Nelson (1990), and has been used for volatility forecasting. Another related representation is the square-root process which belongs to the affine model class and allows for analytically tractable pricing of derivatives, as discussed in more detail later. Wiggins (1987) also starts from the general univariate diffusion (9) but then focuses on the special case where log volatility follows a Gaussian Ornstein-Uhlenbeck (OU) process,

$$\mathrm{d}\log\sigma^2 = \alpha(\mu - \log\sigma^2)\mathrm{d}t + \omega\mathrm{d}B, \quad \alpha > 0. \tag{11}$$

The log-normal SV model of Taylor (1982) can be thought of as an Euler discretization to this continuous-time model over a unit time period. Ito's formula implies that this log-normal OU model can be written as

$$\mathrm{d}\sigma^2 = \left\{\theta - \alpha \log \sigma^2\right\} \sigma^2 \mathrm{d}t + \omega \sigma^2 \mathrm{d}B. \tag{12}$$

It is evident that it resembles the previous models in important respects although it is also distinctly different in the drift specification.

The initial diffusion-based SV models specify volatility to be Markovian with continuous sample paths. This is a constraint on the general SV structure (8) which requires neither of these assumptions. Research in the late 1990s and early 2000s has shown that more complex volatility dynamics are needed to model either options data or high-frequency return data. Leading extensions to the model are to allow jumps in the volatility SDE, e.g., Barndorff-Nielsen and Shephard (2001) and Eraker et al. (2003) or to model the volatility process as a function of a number of separate stochastic processes or factors, e.g., Chernov et al. (2003), Barndorff-Nielsen and Shephard (2001).

A final noteworthy observation is that SV models and time-changed Brownian motions provide fundamental representations for continuous-time martingales. If M is a process with continuous martingale sample paths then the celebrated Dambis-Dubins-Schwartz Theorem, e.g., Rogers and Williams (1996, p. 64), ensures that M can be written as a time-changed BM with the time-change being the quadratic variation (QV) process ,

$$[M]_t = \mathrm{p}\text{-}\lim \sum_{j=1}^{n} \left(M_{t_j} - M_{t_j-1}\right)^2, \tag{13}$$

for any sequence of partitions $t_0 = 0 < t_1 < ... < t_n = t$ with $\sup_j \{t_j - t_{j-1}\} \to 0$ for $n \to \infty$. What is more, as M has continuous sample paths, so must $[M]$. Under the stronger condition that $[M]$ is absolutely continuous, M can be written as a stochastic volatility process. This latter result, known as the martingale representation theorem, is due to Doob (1953). Taken together this implies that time-changed BMs are canonical in continuous sample path price processes and SV models arise as special cases. In the SV case we thus have,

$$[M]_t = \int_0^t \sigma_s^2 \mathrm{d}s. \tag{14}$$

Hence, the increments to the quadratic variation process are identical to the corresponding integrated return variance generated by the SV model.

3 Second Generation Model Building

3.1 Univariate models

3.1.1 Jumps

All the work discussed previously assumes that the asset price process is continuous. Yet, theory asserts that discrete changes in price should occur when significant new information is revealed. In fact, equity indices, Treasury bonds and foreign exchange rates all do appear to jump at the moment significant macroeconomic or monetary policy news are announced. Likewise, individual stock prices often react abruptly to significant company-specific news like earnings reports, see, e.g. Andersen et al. (2007) and Johannes and Dubinsky (2006). As long as these jumps are unknown in terms of timing and/or magnitude this remains consistent with the no-arbitrage semimartingale setting subject only to weak regularity conditions. The cumulative sum of squared price jumps contribute to the return quadratic variation, thus generating distinct diffusive (integrated variance) and jump components in volatility.

Moreover, empirical work using standard SV models, extended by adding jumps to the price process, document significant improvements in model fit, e.g., Andersen et al. (2002) and Eraker et al. (2003). This follows, of course, earlier theoretical work by Merton (1976) on adding jumps to the Black-Scholes diffusion. Bates (1996) was particularly important for the option pricing literature as he documents the need to include jumps in addition to SV for derivatives pricing, at least when volatility is Markovian.

Another restrictive feature of the early literature was the absence of jumps in the diffusive volatility process. Such jumps are considered by Eraker et al. (2003) who deem this extension critical for adequate model fit. A very different approach for SV models was put forth by Barndorff-Nielsen and Shephard (2001) who build volatility models from pure jump processes. In particular, in their simplest model, σ^2 represent the solution to the SDE

$$\mathrm{d}\sigma_t^2 = -\lambda\sigma_t^2\mathrm{d}t + \mathrm{d}z_{\lambda t}, \quad \lambda > 0, \tag{15}$$

where z is a subordinator with independent, stationary and non-negative increments. The unusual timing convention for $z_{\lambda t}$ ensures that the stationary distribution of σ^2 does not depend on λ. These non-Gaussian OU processes are analytically tractable as they belong to the affine model class discussed below.

Geman et al. (2002) provide a new perspective within the general setting by defining the martingale component of prices as a time-change Lévy process, generalizing Clark's time-change of Brownian motion. Empirical evidence in Barndorff-Nielsen and Shephard (2006) suggest these rather simple models

may potentially perform well in practice. Note, if one builds the time-change of the pure jump Lévy process from of an integrated non-Gaussian OU process then the resulting process will not have any Brownian components in the continuous-time price process.

3.1.2 Long memory

In the first generation of SV models the volatility process was given by a simple SDE driven by a BM. This implies that spot volatility is a Markov process. There is considerable empirical evidence that, whether volatility is measured using high-frequency data over a few years or using daily data recorded over decades, the dependence in the volatility structure decays at a rapid rate for shorter lags, but then at a much slower hyperbolic rate at longer lags. Moreover, consistent with the hypothesis that long memory is operative in the volatility process, the estimates for the degree of fractional integration appear remarkably stable irrespective of the sampling frequencies of the underlying returns or the sample period, see Andersen and Bollerslev (1997). As an alternative, it is possible to approximate the long memory feature well by specifying the (log) volatility process via a sum of first-order autoregressive components, leading to multi-factor SV models as pursued by, e.g., Chernov et al. (2003).

The literature has been successful in directly accommodating the longer run volatility dependencies through both discrete-time and continuous-time long memory SV models. In principle, this is straightforward as it only requires specifying a long-memory model for σ. Breidt et al. (1998) and Harvey (1998) study discrete-time models where log volatility is modeled as a fractionally integrated process. They show this can be handled econometrically by quasi-likelihood estimators which are computationally simple, although not fully efficient. In continuous time Comte and Renault (1998) model log volatility as a fractionally integrated BM. More recent work includes the infinite superposition of non-negative OU processes introduced by Barndorff-Nielsen (2001). The two latter models have the potential advantage that they can be used for options pricing without excessive computational effort.

3.2 Multivariate models

Diebold and Nerlove (1989) cast a multivariate SV model within the factor structure used in many areas of asset pricing. Restated in continuous time, their model for the $(N \times 1)$ vector of martingale components of the log asset price vector takes the form,

$$M = \sum_{j=1}^{J} \left(\beta_{(j)} F_{(j)}\right) + G, \tag{16}$$

where the factors $F_{(1)}$,$F_{(2)}$,...,$F_{(J)}$ are independent univariate SV models, $J < N$, and G is a correlated $(N \times 1)$ BM, and the $(N \times 1)$ vector of factor loadings, $\beta_{(j)}$, remains constant through time. This structure has the advantage that the martingale component of time-invariant portfolios assembled from such assets will inherit this basic factor structure. Related papers on the econometrics of this model structure and their empirical performance include King et al. (1994) and Fiorentini et al. (2004).

A more limited multivariate discrete-time model was put forth by Harvey et al. (1994) who suggest having the martingale components be given as a direct rotation of a p-dimensional vector of univariate SV processes. Another early contribution was a multivariate extension of Jacquier et al. (1994) which evolved into Jacquier et al. (1999). In recent years, the area has seen a dramatic increase in activity as is evident from the chapter on multivariate SV in this handbook, cf. Chib et al. (2008).

4 Inference Based on Return Data

4.1 Moment–based inference

A long standing difficulty for applications based on SV models was that the models were hard to estimate efficiently in comparison with their ARCH cousins due to the latency of the volatility state variable. In ARCH models, by construction, the likelihood (or quasi-likelihood) function is readily available. In SV models this is not the case which early on inspired two separate approaches. First, there is a literature on computationally intensive methods which approximate the efficiency of likelihood-based inference arbitrarily well, but at the cost of using specialized and time-consuming techniques. Second, a large number of papers have built relatively simple, inefficient estimators based on easily computable moments of the model. We briefly review the second literature before focusing on the former. We will look at the simplification high frequency data brings to these questions in Section 6.

The task is to carry out inference based on a sequence of returns $y = (y_1, ..., y_T)'$ from which we will attempt to learn about $\theta = (\theta_1, ..., \theta_K)'$, the parameters of the SV model. The early SV paper by Taylor (1982) calibrated the discrete-time model using the method of moments. Melino and Turnbull (1990) improve the inference by relying on a larger set of moment conditions and combining them more efficiently as they exploit the generalized method of moments (GMM) procedure. The quality of the (finite sample) GMM inference is quite sensitive to both the choice of the number of moments to include

and the exact choice of moments among the natural candidates. Andersen and Sørensen (1996) provide practical guidelines for the GMM implementation and illustrate the potentially sizeable efficiency gains in the context of the discrete-time lognormal SV model. One practical drawback is that a second inference step is needed to conduct inference regarding the realizations of the latent volatility process. A feasible approach is to use a linear Kalman filter approximation to the system, given the first stage point estimates for the parameters, and extract the volatility series from the filter. However, this is highly inefficient and the combination of a two-step approach and a relatively crude approximation renders it hard to assess the precision of the inference for volatility.

Harvey et al. (1994) apply the natural idea of using the Kalman filter for joint quasi-likelihood estimation of the model parameters and the time-varying volatility for the log-normal SV model defined via (5) and (7). This method produces filtered as well as smoothed estimates of the underlying volatility process. The main drawback is that the method is quite inefficient as the linearized system is highly non-Gaussian.

For continuous-time SV models, it is generally much harder to derive the requisite closed form solutions for the return moments. Nonetheless, Meddahi (2001) provides a general approach for generating moment conditions for the full range of models that fall within the so-called Eigenfunction SV class. A thorough account of the extensive literature on moment-based SV model inference, including simulation-based techniques, is given in Renault (2008).

4.2 Simulation–based inference

Within the last two decades, a number of scholars have started to develop and apply simulation-based inference devices to tackle SV models. Concurrently two approaches were brought forward. The first was the application of Markov chain Monte Carlo (MCMC) techniques. The second was the development of indirect inference or the so-called efficient method of moments. To discuss these methods it is convenient to focus on the simplest discrete-time log-normal SV model given by (5) and (7).

MCMC allows us to simulate from high dimensional posterior densities, such as the smoothing variables $h|y, \theta$, where $h = (h_1, ..., h_T)'$ are the discrete time unobserved log-volatilities. Shephard (1993) notes that SV models are a special case of a Markov random field so MCMC can be used for simulation of $h|y, \theta$. Hence, the simulation output inside an EM algorithm can be used to approximate the maximum likelihood estimator of θ. However, the procedure converges slowly. Jacquier et al. (1994) demonstrate that a more elegant inference may be developed by becoming Bayesian and using the MCMC algorithm to simulate from $h, \theta|y$. Once the ability to compute many simulations from this $T + K$ dimensional random variable (there are K pa-

rameters), one can discard the h variables and simply record the many draws from $\theta|y$. Summarizing these draws allows for fully efficient parametric inference in a relatively sleek way. Later, Kim et al. (1998) provide an extensive discussion of alternative methods for implementing the MCMC algorithm. This is a subtle issue and can make a large difference to the computational efficiency of the methods.

Kim et al. (1998) also introduce a genuine filtering method for recursively sampling from

$$h_1, ..., h_i|y_1, ..., y_{i-1}, \theta, \quad i = 1, 2, ..., T. \tag{17}$$

These draws enable estimation, by simulation, of $\mathrm{E}\left(\sigma_i^2|y_1, ..., y_{t-1}, \theta\right)$ as well as the corresponding density and the density of $y_i|y_1, ..., y_{t-1}, \theta$ using the so-called particle filter, see, e.g., Gordon et al. (1993) and Pitt and Shephard (1999). These quantities are useful inputs for financial decision making as they are derived conditionally only on current information. Moreover, they allow for computation of marginal likelihoods for model comparison and for one-step-ahead predictions for specification testing. [3] Although these MCMC based papers are couched in discrete time, it is also noted that the general approach can be adapted to handle models operating with data generated at higher frequencies through data augmentation. This strategy was implemented for diffusion estimation by Jones (1998), Eraker (2001), Elerian et al. (2001), and Roberts and Stramer (2001).

The MCMC approach works effectively under quite general circumstances, although it is dependent on the ability to generate appropriate and efficient proposal densities for the potentially complex conditional densities that arise during the recursive sampling procedure. An alternative is to develop a method that maximizes a simulation based estimate of the likelihood function. This may require some case-by-case development but it has been implemented for a class of important discrete-time models by Danielsson and Richard (1993) using the Accelerated Gaussian Importance Sampler. The procedure was further improved through improved simulation strategies by Fridman and Harris (1998) and Leisenfeld and Richard (2003). A formal approach for simulated maximum likelihood estimation of diffusions is developed by Pedersen (1995) and simultaneously, with a more practical orientation, by Santa-Clara (1995). Later refinements and applications for SV diffusion models include Elerian et al. (2001), Brandt and Santa-Clara (2002), Durham and Gallant (2002), and Durham (2003).

Another successful approach for diffusion estimation was developed via a novel extension to the Simulated Method of Moments of Duffie and Singleton (1993). Gouriéroux et al. (1993) and Gallant and Tauchen (1996) propose to fit the moments of a discrete-time auxiliary model via simulations from the underlying continuous-time model of interest, thus developing the approach into what is now termed Indirect Inference or the Efficient Method of Mo-

[3] A detailed account of the particle filter is given by Johannes and Polson in this handbook, cf. Johannes and Polson (2008).

ments (EMM). The latter approach may be intuitively explained as follows. First, an auxiliary model is chosen to have a tractable likelihood function but with a generous parameterization that should ensure a good fit to all significant features of the time series at hand. For financial data this typically involves an ARMA-GARCH specification along with a dynamic and richly parameterized (semi-nonparametric or SNP) representation of the density function for the return innovation distribution. The auxiliary model is estimated by (quasi-) likelihood from the discretely observed data. This provides a set of score moment functions which, ideally, encode important information regarding the probabilistic structure of the actual data sample. Next, a very long sample is simulated from the continuous-time model. The underlying continuous-time parameters are varied in order to produce the best possible fit to the quasi-score moment functions evaluated on the simulated data. If the underlying continuous-time model is correctly specified it should be able to reproduce the main features of the auxiliary score function extracted from the actual data. It can be shown, under appropriate regularity, that the method provides asymptotically efficient inference for the continuous-time parameter vector. A useful side-product is an extensive set of model diagnostics and an explicit metric for measuring the extent of failure of models which do not adequately fit the quasi-score moment function. Gallant et al. (1997) provide an in-depth discussion and illustration of the use of these methods in practice. Moreover, the task of forecasting volatility conditional on the past observed data (akin to filtering in MCMC) or extracting volatility given the full data series (akin to smoothing in MCMC) may be undertaken in the EMM setting through the reprojection method developed and illustrated in Gallant and Tauchen (1998).

An early use of Indirect Inference for SV diffusion estimation is Engle and Lee (1996) while EMM has been extensively applied with early work exploring short rate volatility (Andersen and Lund (1997)), option pricing under SV (Chernov and Ghysels (2000)), affine and quadratic term structure models (Dai and Singleton (2000), Ahn et al. (2002)), SV jump-diffusions for equity returns (Andersen et al. (2002)) and term structure models with regime-shifts (Bansal and Zhou (2002)).

An alternative approach to estimation of spot volatility in continuous time is given by Foster and Nelson (1996). They develop an asymptotic distribution theory for a local variance estimator, computed from the lagged data,

$$\widehat{\sigma_t^2} = h^{-1} \sum_{j=1}^{M} \left(Y_{t-hj/M} - Y_{t-h(j-1)/M} \right)^2 . \tag{18}$$

They study the behavior of the estimator as $M \to \infty$ and $h \downarrow 0$ under a set of regularity conditions, ruling out, e.g., jumps in price or volatility. This "double asymptotics" yields a Gaussian limit theory as long as $h \downarrow 0$ and $M \to \infty$ at the correct, connected rates. This is related to the realized volatility approach detailed in a separate section below although, importantly, the latter focuses

on the integrated volatility rather than the spot volatility and thus avoids some of the implementation issues associated with the double limit theory.

5 Options

5.1 Models

As discussed previously, the main impetus behind the early SV diffusion models was the desire to obtain a realistic basis for option pricing. A particularly influential contribution was Hull and White (1987) who studied a diffusion with leverage effects. Assuming volatility risk is fully diversifiable, they price options either by approximation or by simulation. The results suggest that SV models are capable of producing smiles and skews in option implied volatilities as often observed in market data. Renault (1997) studies these features systematically and confirms that smiles and smirks emerge naturally from SV models via leverage effects.

The first analytic SV option pricing formula is by Stein and Stein (1991) who model σ as a Gaussian OU process. European option prices may then be computed using a single Fourier inverse which, in this literature, is deemed "closed form." A conceptual issue with the Gaussian OU model is that it allows for a negative volatility process. Heston (1993) overcomes this by employing a version of the so-called square root volatility process. Bates (1996) extends the framework further to allow for jumps in the underlying price and shows that these are critical for generating a reasonable fit to option prices simultaneously across the strike and time-to-maturity spectrum. Another closed-form option pricing solution is given by Nicolato and Venardos (2003) who rely on the non-Gaussian OU SV models of Barndorff-Nielsen and Shephard (2001).

All models above belong to the affine class advocated by Duffie et al. (2000). These models are used extensively because they provide analytically tractable solutions for pricing a wide range of derivative securities. The general case involves solving a set of ordinary differential equations inside a numerical Fourier inverse but this may be done quickly on modern computers. These developments have spurred more ambitious inference procedures for which the parameters of affine SV models for both the underlying asset and the risk-neutral dynamics governing market pricing are estimated jointly from data on options and the underlying. Chernov and Ghysels (2000) estimate the affine SV diffusions for the actual and risk-neutral measures simultaneously using EMM. Pan (2002) exploits at-the-money options while allowing for an affine SV jump-diffusion representation under the actual and risk-neutral measure. Her inference is conducted via GMM, exploiting the closed-form expressions for the joint conditional moment-generating function

of stock returns and volatility developed in Duffie et al. (2000); see also Singleton (2001). Eraker (2004) expands the model specification, using MCMC based inference, to include a wider cross-section of option strikes and allowing for jumps in the volatility process as well. Finally, it is possible to develop option pricing on time-change Lévy processes, see, e.g., Carr and Wu (2004) who develop the derivatives pricing in a setting inspired by Geman et al. (2002).

6 Realized Volatility

A couple of relatively recent developments have moved SV models towards the center of volatility research. This process is related to the rapid increase in research under the general heading of realized volatility.

One major change is the advent of commonly available and very informative high-frequency data, such as minute-by-minute return data or entire records of quote and/or transaction price data for particular financial instruments. The first widely disseminated data of this type were foreign exchange quotes gathered by Olsen & Associates, discussed in detail in the seminal work of Dacorogna et al. (2001). Later scholars started using tick-by-tick data from the main equity and futures exchanges in the U.S. and Europe. This naturally moved the perspective away from fixed time intervals, such as a day, and into the realm where, at least in theory, one thinks of inference regarding the price process over different horizons based on ever changing information sets. This type of analysis is, of course, ideally suited to a continuous-time setting as any finite-horizon distribution then, in principle, may be obtained through time aggregation. Moreover, this automatically ensures modeling coherence across different sampling frequencies. Hence, almost by construction, volatility clustering in continuous time points us towards SV models.

A related development is the rapidly accumulating theoretical and empirical research on how to exploit this high-frequency data to estimate the increments of the quadratic variation (QV) process and then to use this estimate to project QV into the future in order to predict future levels of volatility. This literature deals with various aspects of so-called realized variation, also often more generically referred to as realized volatility. This section briefly introduces some of the main ideas, leaning on contributions from Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002). A more detailed account is given in the chapter by Andersen and Benzoni in this handbook (cf. Andersen and Benzoni (2008)).

In realized variation theory, high-frequency data are used to estimate the QV process. We let δ denote a time period between high-frequency observations and define the *realized QV process* as,

$$[Y_\delta]_t = \sum_{j=1}^{\lfloor t/\delta \rfloor} \{Y(\delta j) - Y(\delta (j-1))\}^2 . \tag{19}$$

Then, by the definition of the QV process, as $\delta \downarrow 0$,

$$[Y_\delta]_t \xrightarrow{p} [Y]_t, \tag{20}$$

which the probability literature has shown to be well behaved if Y is a semimartingale. If the expected return process has continuous sample paths, then $[Y] = [M]$, and if additionally M is a SV process then $[Y_\delta]_t \xrightarrow{p} \int_0^t \sigma_s^2 \mathrm{d}s$.

In practice, it is preferable to measure increments of the quadratic variation process over one full trading day (or week). This measure is often referred to as the daily realized variance while its square root then is denoted the daily realized volatility, following the terminology of the financial mathematics literature. This should not be confused with the more generic terminology that refers to all transformations of realized quadratic variation measures as realized volatility. The main reason for aggregating the realized variation measures to a daily frequency is the presence of pronounced and systematic intraday patterns in return volatility. These stem from highly regular, but dramatic, shifts in the quote and transactions intensity across the trading day as well as the release of macroeconomic and financial news according to specific time tables. Often, new information creates short-run dynamics akin to a price discovery process with an immediate price jump followed by a brief burst in volatility, see, e.g., Andersen and Bollerslev (1998). As a result, the intraday volatility process displays rather extreme variation and contains various components with decidedly low volatility persistence. Consequently, the direct modeling of the ultra high-frequency volatility process is both complex and cumbersome. Yet, once the return variation process is aggregated into a time series of daily increments, the strong inter-daily dependence in return volatility is brought out very clearly as the systematic intraday variation, to a large extent, is annihilated by aggregation across the trading day. In fact, the evidence for inter-daily volatility persistence is particularly transparent from realized volatility series compared to the traditional volatility measures inferred from daily return data.

Andersen et al. (2001) show that a key input for forecasting the volatility of future asset returns should be predictions of the future daily quadratic return variation. Recall from Ito's formula that, if Y is a continuous sample path semimartingale then

$$Y_t^2 = [Y]_t + 2\int_0^t Y_s \mathrm{d}Y_s = [Y]_t + 2\int_0^t Y_s \mathrm{d}A_s + 2\int_0^t Y_s \mathrm{d}M_s. \tag{21}$$

Letting $\mathcal{F}_t$ denote the filtration generated by the continuous history of Y_t up to time t and exploiting that M is a martingale, we have

$$\mathrm{E}(Y_t^2|\mathcal{F}_0) = \mathrm{E}\left([Y]_t|\mathcal{F}_0\right) + 2\mathrm{E}\left(\int_0^t Y_s \mathrm{d}A_s|\mathcal{F}_0\right). \tag{22}$$

In practice, over small intervals of time, the second term is small, so that

$$\mathrm{E}(Y_t^2|\mathcal{F}_0) \simeq \mathrm{E}\left([Y]_t|\mathcal{F}_0\right). \tag{23}$$

This implies that forecasting future squared daily returns can be done effectively through forecasts for future realized QV increments. A natural procedure estimates a time series model directly from the past observable realized daily return variation and uses it to generate predictions for future realized variances, as implemented through an ARFIMA model for realized log volatility in Andersen et al. (2003). The incorporation of long memory through fractional integration proves particularly important for forecast performance while only a few autoregressive lags are needed to accommodate shorter run dependencies. Hence, long lags of appropriately weighted (hyperbolically decaying) realized log volatilities prove successful in forecasting future volatility.

A potential concern with this approach is that the QV theory only tells us that $[Y_\delta] \xrightarrow{p} [Y]$, but does not convey information regarding the likely size of the measurement error, $[Y_\delta]_t - [Y]_t$. Jacod (1994) and Barndorff-Nielsen and Shephard (2002) strengthen the consistency result to provide a central limit theory for the univariate version of this object. They show that the measurement errors are asymptotically uncorrelated and

$$\frac{\delta^{-1/2}\left([Y_\delta]_t - [Y]_t\right)}{\sqrt{2\int_0^t \sigma_s^4 \mathrm{d}s}} \xrightarrow{d} N(0,1). \tag{24}$$

The latter also develop a method for consistently estimating the integrated quarticity, $\int_0^t \sigma_s^4 \mathrm{d}s$, from high-frequency data, thus enabling feasible inference on the basis of the above result. This analysis may help simplify parametric estimation as we obtain estimates of the key volatility quantities that SV models directly parameterize. In terms of volatility forecasting, the use of long lags of weighted realized volatilities tends to effectively diversify away the impact of measurement errors so that the predictive performance is less adversely impacted than one may suspect, see Andersen et al. (2006).

In the very recent past there have been various elaborations to this literature. We briefly mention two. First, there has been interest in studying the impact of market microstructure effects on the estimates of realized variance. This causes the estimator of the QV to become biased. Leading papers on this topic are Hansen and Lunde (2006), Zhang et al. (2005), Bandi and Russell (2006) and Barndorff-Nielsen et al. (2008). Second, one can estimate the QV of the continuous component of prices in the presence of jumps using the so-called realized bipower variation process. This was introduced by Barndorff-Nielsen and Shephard (2004).

References

Ahn, D.-H., Dittmar, R. F. and Gallant, A. R. (2002): Quadratic term structure models: Theory and evidence. *Review of Financial Studies* **15**, 243–288.

Andersen, T. G. (1996): Return volatility and trading volume: an information flow interpretation of stochastic volatility. *Journal of Finance* **51**, 169–204.

Andersen, T. G., Benzoni, L. and Lund, J. (2002): An empirical investigation of continuous-time equity return models. *Journal of Finance* **57**, 1239–1284.

Andersen, T. G. and Benzoni, L. (2008): Realized volatility. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 554–575. Springer, New York.

Andersen, T. G. and Bollerslev, T. (1997): Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *Journal of Finance* **52**, 975–1005.

Andersen, T. G. and Bollerslev, T. (1998): Deutsche mark-dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* **53**, 219–265.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001): The distribution of exchange rate volatility. *Journal of the American Statistical Association* **96**, 42–55. Correction published in 2003, volume 98, page 501.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003): Modeling and forecasting realized volatility. *Econometrica* **71**, 579–625.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Vega, C. (2007): Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics* **73**, 251–277.

Andersen, T. G., Bollerslev, T. and Meddahi, N. (2006): Market microstructure noise and realized volatility forecasting. *Unpublished paper. Department of Economics, Duke University*.

Andersen, T. G. and Lund, J. (1997): Estimating continuous-time stochastic volatility models of the short term interest rate. *Journal of Econometrics* **2**, 343–77.

Andersen, T. G. and Sørensen, B. (1996): GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Business and Economic Statistics* **14**, 328–352.

Back, K. (1991): Asset pricing for general processes. *Journal of Mathematical Economics* **20**, 371–395.

Bandi, F. M. and Russell, J. R. (2006): Seperating microstructure noise from volatility. *Journal of Financial Economics* **79**, 655–692.

Bansal, R. and Zhou, H. (2002): Term structure of interest rates with regime shifts. *Journal of Finance* **57**, 1997–2043.

Barndorff-Nielsen, O. E. (2001). Superposition of Ornstein-Uhlenbeck type processes. *Theory of Probability and its Applications* **46**, 175–194.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008): Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, forthcoming.

Barndorff-Nielsen, O. E. and Shephard, N. (2001): Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, Series B* **63**, 167–241.

Barndorff-Nielsen, O. E. and Shephard, N. (2002): Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* **64**, 253–280.

Barndorff-Nielsen, O. E. and Shephard, N. (2004): Power and bipower variation with stochastic volatility and jumps (with discussion). *Journal of Financial Econometrics* **2**, 1–48.

Barndorff-Nielsen, O. E. and Shephard, N. (2006): Impact of jumps on returns and realised variances: econometric analysis of time-deformed Lévy processes. *Journal of Econometrics* **131**, 217–252.

Bates, D. S. (1996): Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of Financial Studies* **9**, 69–107.

Black, F. (1976): Studies of stock price volatility changes. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 177–181.

Black, F. and Scholes, M. (1972): The valuation of option contracts and a test of market efficiency. *Journal of Finance* **27**, 399–418.

Brandt, M. W. and Santa-Clara, P. (2002): Simulated likelihood estimation of diffusions with an application to exchange rates dynamics in incomplete markets. *Journal of Financial Economics* **63**, 161–210.

Breidt, F. J., Crato, N. and de Lima, P. (1998): On the detection and estimation of long memory in stochastic volatility. *Journal of Econometrics* **83**, 325–348.

Carr, P. and Wu, L. (2004): Time-changed Lévy processes and option pricing. *Journal of Financial Economics* **71**, 113–141.

Chernov, M., Gallant, A. R., Ghysels, E. and Tauchen, G. (2003): Alternative models of stock price dynamics. *Journal of Econometrics* **116**, 225–257.

Chernov, M. and Ghysels, E. (2000): A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of Financial Economics* **56**, 407–458.

Chib, S., Omori, Y. and Asai, M. (2008): Multivariate stochastic volatility. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 365–400. Springer, New York.

Clark, P. K. (1973): A subordinated stochastic process model with fixed variance for speculative prices. *Econometrica* **41**, 135–156.

Comte, F. and Renault, E. (1998): Long memory in continuous-time stochastic volatility models. *Mathematical Finance* **8**, 291–323.

Dacorogna, M. M., Gencay, R., Müller, U. A., Olsen, R. B. and Pictet, O. V. (2001): *An Introduction to High-Frequency Finance*. Academic Press, San Diego.

Dai, Q. and Singleton, K. J. (2000): Specification analysis of affine term structure models. *Journal of Finance* **55**, 1943–1978.

Danielsson, J. and Richard, J. F. (1993): Accelerated Gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics* **8**, 153–174.

Diebold, F. X. and Nerlove, M. (1989): The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *Journal of Applied Econometrics* **4**, 1–21.

Doob, J. L. (1953): *Stochastic Processes*. John Wiley and Sons, New York.

Duffie, D., Pan, J. and Singleton, K. J. (2000): Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68**, 1343–1376.

Duffie, D. and Singleton, K. J. (1993): Simulated moments estimation of markov models of asset prices. *Econometrica* **61**, 929–952.

Durham, G. (2003): Likelihood-based specification analysis of continuous-time models of the short-term interest rate. *Journal of Financial Economics* **70**, 463–487.

Durham, G. and Gallant, A. R. (2002): Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes (with discussion). *Journal of Business and Economic Statistics* **20**, 297–338.

Elerian, O., Chib, S. and Shephard, N. (2001): Likelihood inference for discretely observed non-linear diffusions. *Econometrica* **69**, 959–993.

Engle, R. F. and Lee, G. G. J. (1996): Estimating diffusion models of stochastic volatility. *In: Rossi, P. E. (Ed.): Modelling Stock Market Volatility – Bridging the GAP to Continuous Time*, 333–384. Academic Press.

Epps, T. W. and Epps, M. L. (1976): The stochastic dependence of security price changes and transaction volumes: implications for the mixture-of-distributions hypothesis. *Econometrica* **44**, 305–321.

Eraker, B. (2001): Markov chain Monte Carlo analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics* **19**, 177–191.

Eraker, B. (2004): Do stock prices and volatility jump? reconciling evidence from spot and option prices. *Journal of Finance* **59**, 1367–1403.

Eraker, B., Johannes, M. and Polson, N. G. (2003): The impact of jumps in returns and volatility. *Journal of Finance* **53**, 1269–1300.

Fama, E. F. (1965): The behaviour of stock market prices. *Journal of Business* **38**, 34–105.

Fiorentini, G., Sentana, E. and Shephard, N. (2004): Likelihood-based estimation of latent generalised ARCH structures. *Econometrica* **72**, 1481–1517.

Foster, D. P. and Nelson, D. B. (1996): Continuous record asymptotics for rolling sample variance estimators. *Econometrica* **64**, 139–174.

Fridman, M. and Harris, L. (1998): A maximum likelihood approach for non-Gaussian stochastic volatility models. *Journal of Business and Economic Statistics* **16**, 284–291.

Gallant, A. R., Hsieh, D. and Tauchen, G. (1997): Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* **81**, 159–192.

Gallant, A. R. and Tauchen, G. (1996): Which moments to match. *Econometric Theory* **12**, 657–81.

Gallant, A. R. and Tauchen, G. (1998): Reprojection partially observed systems with applications to interest rate diffusions. *Journal of the American Statistical Association* **93**, 10–24.

Geman, H., Madan, D. B. and Yor, M. (2002): Stochastic volatility, jumps and hidden time changes. *Finance and Stochastics* **6**, 63–90.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993): A novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE-Proceedings F* **140**, 107–113.

Gouriéroux, C., Monfort, A. and Renault, E. (1993): Indirect inference. *Journal of Applied Econometrics* **8**, S85–S118.

Hansen, P. R. and Lunde, A. (2006): Realized variance and market microstructure noise (with discussion). *Journal of Business and Economic Statistics* **24**, 127–218.

Harvey, A. C. (1998). Long memory in stochastic volatility. *In: Knight, J. and Satchell, S. (Eds.): Forecasting Volatility in Financial Markets*, 307–320. Butterworth-Heineman, Oxford.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994): Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–264.

Heston, S. L. (1993): A closed-form solution for options with stochastic volatility, with applications to bond and currency options. *Review of Financial Studies* **6**, 327–343.

Hull, J. and White, A. (1987): The pricing of options on assets with stochastic volatilities. *Journal of Finance* **42**, 281–300.

Jacod, J. (1994): Limit of random measures associated with the increments of a Brownian semimartingale. *Preprint number 120, Laboratoire de Probabilitiés, Université Pierre et Marie Curie, Paris.*

Jacquier, E., Polson, N. G. and Rossi, P. E. (1994): Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business and Economic Statistics* **12**, 371–417.

Jacquier, E., Polson, N. G. and Rossi, P. E. (1999): Stochastic volatility: Univariate and multivariate extensions. *Unpublished paper: GSB, University of Chicago and HEC, Montreal.*

Johannes, M. and Dubinsky, A. (2006): Earnings announcements and equity options. *Unpublished paper: Columbia University.*

Johannes, M. and Polson, N. (2008): Particle filtering. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 1014– 1029. Springer, New York.

Johnson, H. (1979): Option pricing when the variance rate is changing. *Working paper, University of California, Los Angeles.*

Johnson, H. and Shanno, D. (1987): Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis* **22**, 143–151.

Jones, C. S. (1998): Bayesian estimation of continuous-time finance models. *Working Paper, Simon Graduate School of Business, University of Rochester.*

Kim, S., Shephard, N. and Chib, S. (1998): Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**, 361–393.

King, M., Sentana, E. and Wadhwani, S. (1994): Volatility and links between national stock markets. *Econometrica* **62**, 901–933.

Leisenfeld, R. (2001): A generalized bivariate mixture model for stock price volatility and trading volume. *Journal of Econometrics* **104**, 141–178.

Leisenfeld, R. and Richard, J.-F. (2003): Univariate and multivairate stochastic volatility models: Estimation and diagnostics. *Journal of Empirical Finance* **10**, 505–531.

Mandelbrot, B. (1963): The variation of certain speculative prices. *Journal of Business* **36**, 394–419.

Meddahi, N. (2001): An eigenfunction approach for volatility modeling. *Unpublished paper, University of Montreal.*

Melino, A. and Turnbull, S. M. (1990): Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* **45**, 239–265.

Merton, R. C. (1976): Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* **3**, 125–144.

Nelson, D. B. (1990): ARCH models as diffusion approximations. *Journal of Econometrics* **45**, 7–38.

Nicolato, E. and Venardos, E. (2003): Option pricing in stochastic volatility models of the Ornstein-Uhlenbeck type. *Mathematical Finance* **13**, 445–466.

Pan, J. (2002): The jump-risk premia implicit in option prices: evidence from an integrated time-series study. *Journal of Financial Economics* **63**, 3–50.

Pedersen, A. R. (1995): A new approach to maximum likelihood estimation for stochastic differential equations on discrete observations. *Scandinavian Journal of Statistics* **27**, 55–71.

Pitt, M. K. and Shephard, N. (1999): Filtering via simulation: auxiliary particle filter. *Journal of the American Statistical Association* **94**, 590–599.

Renault, E. (1997): Econometric models of option pricing errors. *In: Kreps, D. M. and Wallis, K. F. (Eds.): Advances in Economics and Econometrics: Theory and Applications*, 23–78. Cambridge University Press.

Renault, E. (2008): Moment-based estimation of stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 268– 311. Springer, New York.

Roberts, G. O. and Stramer, O. (2001): On inference for nonlinear diffusion models using the Hastings-Metropolis algorithms. *Biometrika* **88**, 603–621.

Rogers, L. C. G. and Williams, D. (1996): *Diffusions, Markov Processes and Martingales. Volume 2, Itô Calculus* (2 ed.). Wiley, Chichester.

Rosenberg, B. (1972): The behaviour of random variables with nonstationary variance and the distribution of security prices. *Working paper 11, Graduate School of Business Administration, University of California, Berkeley.* Reprinted in Shephard (2005).

Santa-Clara, P. (1995): *Simulated likelihood estimation of diffusions with an application to the short term interest rate.* PhD Dissertation, INSEAD.

Shephard, N. (1993): Fitting non-linear time series models, with applications to stochastic variance models. *Journal of Applied Econometrics* **8**, S135–52.

Shephard, N. (2005): *Stochastic Volatility: Selected Readings.* Oxford University Press.

Singleton, K. J. (2001): Estimation of affine asset pricing models using the empirical characteristic function. *Journal of Econometrics* **102**, 111–141.

Stein, E. M. and Stein, J. (1991): Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies* **4**, 727–752.

Tauchen, G. and Pitts, M. (1983): The price variability-volume relationship on speculative markets. *Econometrica* **51**, 485–505.

Taylor, S. J. (1982): Financial returns modelled by the product of two stochastic processes — a study of daily sugar prices 1961-79. *In: Anderson, O. D. (Ed.): Time Series Analysis: Theory and Practice* **1**, 203–226. North-Holland, Amsterdam.

Wiggins, J. B. (1987): Option values under stochastic volatilities. *Journal of Financial Economics* **19**, 351–372.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005): A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* **100**, 1394–1411.

Probabilistic Properties of Stochastic Volatility Models

Richard A. Davis and Thomas Mikosch

Abstract We collect some of the probabilistic properties of a strictly stationary stochastic volatility process. These include properties about mixing, covariances and correlations, moments, and tail behavior. We also study properties of the autocovariance and autocorrelation functions of stochastic volatility processes and its powers as well as the asymptotic theory of the corresponding sample versions of these functions. In comparison with the GARCH model (see Lindner (2008)) the stochastic volatility model has a much simpler probabilistic structure which contributes to its popularity.

1 The Model

We consider a *stochastic volatility process* $(X_t)_{t\in\mathbb{Z}}$ given by the equations

$$X_t = \sigma_t Z_t\,, \quad t \in \mathbb{Z}\,, \tag{1}$$

where $(\sigma_t)_{t\in\mathbb{Z}}$ is a strictly stationary sequence of positive random variables which is independent of the iid *noise* sequence $(Z_t)_{t\in\mathbb{Z}}$.[1] We refer to $(\sigma_t)_{t\in\mathbb{Z}}$

Richard A. Davis
Department of Statistics, 1255 Amsterdam Avenue, Columbia University, New York, NY 10027, U.S.A., www.stat.columbia.edu/∼rdavis, e-mail: rdavis@stat.columbia.edu

Thomas Mikosch
Laboratory of Actuarial Mathematics, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark, www.math.ku.dk/∼mikosch, e-mail: mikosch@math.ku.dk

[1] It is common to assume the additional standardization conditions $EZ_t = 0$ and $\mathrm{var}(Z_t) = 1$. These conditions are important for example in order to avoid identification problems for the parameters of the model. In most parts of this article, these additional conditions are not needed. Moreover, in Sections 4 and 5 we will also consider results when $\mathrm{var}(Z) = \infty$ or $E|Z| = \infty$.

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_11,

as the *volatility* sequence. Following the tradition in time series analysis, we index the stationary sequences (X_t), (Z_t), (σ_t) by the set $\mathbb{Z}$ of the integers. For practical purposes, one would consider e.g., the sequence $(X_t)_{t\in\mathbb{N}}$ corresponding to observations at the times $t = 1, 2, \ldots$.

2 Stationarity, Ergodicity and Strong Mixing

2.1 Strict stationarity

The independence of the noise (Z_t) and the volatility sequence (σ_t) allow for a much simpler probabilistic structure than that of a GARCH process which includes explicit feedback of the current volatility with previous volatilities and observations. This is one of the attractive features of stochastic volatility models. For example, the mutual independence of the sequences (Z_t) and (σ_t) and their strict stationarity immediately imply that (X_t) is strictly stationary.[2] Conditions for the existence of a stationary GARCH process are much more difficult to establish, see Nelson (1990) and Bougerol and Picard (1992), cf. Lindner (2008).

A convenient way of constructing a positive stationary volatility sequence (σ_t) is to assume that $Y_t = \log \sigma_t$, $t \in \mathbb{Z}$, is a stationary sequence (Y_t) with certain nice properties. An obvious candidate is the class of causal linear processes given by

$$Y_t = \sum_{i=0}^{\infty} \psi_i \, \eta_{t-i} \,, \quad t \in \mathbb{Z} \,, \tag{2}$$

where (ψ_t) is a sequence of deterministic coefficients with $\psi_0 = 1$ and (η_t) is an iid sequence of random variables. It immediately follows by an application of Kolmogorov's 3-series theorem (cf. Billingsley (1995), Petrov (1995)) that, if $E\eta = 0$ and $\mathrm{var}(\eta) < \infty$, the infinite series (2) converges a.s. if and only if $\sum_j \psi_j^2 < \infty$. The class of processes (2) includes short memory ARMA as well as long memory fractional ARIMA processes. We refer to Brockwell and Davis (1991) for an extensive treatment of such processes, see also the discussion in Section 3 below. Moreover, if we further specify η to be centered normal and if we assume that $\sum_j \psi_j^2 < \infty$ then the sequence (σ_t) is stationary with log-normal marginals.

In what follows, we always assume that (σ_t), hence (X_t), is a (strictly) stationary sequence.

[2] We refer to stationarity as strict stationarity and we write A for a generic element of any stationary sequence (A_t).

2.2 Ergodicity and strong mixing

If (σ_t) is stationary ergodic, then the sequence (X_t) inherits ergodicity as well. This applies e.g., if $\sigma_t = f((\eta_{t+h})_{h\in\mathbb{Z}})$ for a measurable function f and an iid sequence (η_t), and in particular for the model (2). These facts follow from standard ergodic theory, cf. Krengel (1985).

The process (X_t) also inherits more subtle properties of the volatility sequence such as certain mixing properties. We illustrate this in the case of *strong mixing*. Recall that a stationary sequence $(A_t)_{t\in\mathbb{Z}}$ is *strongly mixing* if it satisfies the property

$$\sup_{B\in\mathcal{F}_{-\infty}^{0}\,,\,C\in\mathcal{F}_{t}^{\infty}} |P(B\cap C) - P(B)\,P(C)| =: \alpha_t \to 0\,, \quad t\to\infty\,,$$

where $\mathcal{F}_a^b$ is the σ-field generated by A_t, $a\le t\le b$, with the obvious modifications for $a=-\infty$ and $b=\infty$. The function (α_t) is called the *mixing rate function* of (A_t). Its decay rate to zero as $t\to\infty$ is a measure of the range of dependence or of the memory in the sequence (A_t). If (α_t) decays to zero at an exponential rate, then (A_t) is said to be strongly mixing with *geometric rate*. In this case, the memory between past and future dies out exponentially fast. A recent survey on strong mixing and its interrelationship with other mixing conditions can be found in Bradley (2005), see also the collection of surveys on dependence Doukhan et al. (2004), Eberlein et al. (1986) and the overviews on mixing properties of time series models in Fan and Yao (2003).

The rate function (α_t) is closely related to the decay of the *autocovariance* and *autocorrelation functions* (ACVF and ACF) of the stationary process (A_t) given by

$$\gamma_A(h) = \mathrm{cov}(A_0, A_h) \quad \text{and} \quad \rho_A(h) = \mathrm{corr}(A_0, A_h)\,, \quad h\ge 0\,, \tag{3}$$

where we assume that $\gamma_A(0) = \mathrm{var}(A) < \infty$. For example, if $E(|A|^{2+\delta}) < \infty$ for some $\delta > 0$, then

$$|\rho_A(h)| \le c\,\alpha_h^{\delta/(2+\delta)} \quad h\ge 0\,, \tag{4}$$

for some constant $c > 0$, see Ibragimov and Linnik (1971), Theorem 17.2.2. In some special cases one can also conclude from the decay rate to zero of the ACVF about the convergence rate to zero of (α_t). For example, if (A_t) is a Gaussian causal ARMA process it is well known (cf. Brockwell and Davis (1991)) that $\rho_A(h)\to 0$ exponentially fast which in turn implies that $\alpha_t\to 0$ exponentially fast; see Pham and Tran (1985).[3]

[3] Pham and Tran (1985) proved the result for β-mixing which implies strong mixing. They also prove the result for general classes of linear processes with exponentially decaying coefficients ψ_i and iid noise (η_i) more general than Gaussian noise.

Since strong mixing is defined via σ-fields, strong mixing of the log-volatility sequence (Y_t) immediately transfers to sequences of measurable functions of the Y_t's. For example, if (Y_t) is strongly mixing with rate function (α_t), so are (σ_t) and (σ_t^2) with the same rate function. Moreover, the stochastic volatility process (X_t) inherits strong mixing from (σ_t) essentially with the same rate. This can be established using the following simple calculation. Since (σ_t) and (Z_t) are independent, we have for any Borel sets $B \in \mathcal{F}_{-\infty}^0$ and $C \in \mathcal{F}_t^\infty$ that

$$
\begin{aligned}
&|\, P(B \cap C) - P(B)\, P(C)| \qquad (5)\\
&= |\, E[f(\ldots, \sigma_{-1}, \sigma_0) g(\sigma_k, \sigma_{k+1}, \ldots)] - E[f(\ldots, \sigma_{-1}, \sigma_0)]\; E[g(\sigma_k, \sigma_{k+1}, \ldots)]|\,,
\end{aligned}
$$

where

$$
\begin{aligned}
f(\ldots, \sigma_{-1}, \sigma_0) &= P\left((\ldots, X_{-1}, X_0) \in A \mid \sigma_s,\, s \le 0\right)\,,\\
g(\sigma_t, \sigma_{t+1}, \ldots) &= P\left((X_t, X_{t+1}, \ldots) \in B \mid \sigma_s,\, s \ge t\right)\,,
\end{aligned}
$$

and standard results about strong mixing (cf. Doukhan (1994)) show that the right-hand side in (5) is bounded by $4\alpha_t$. Finally, we conclude that all sequences (X_t), (σ_t), (σ_t^2), (Y_t) essentially have the same strong mixing properties.

Moreover, a sequence generated by measurable transformations of the form $f(\sigma_t, \ldots, \sigma_{t+h})$ or $g(X_t, \ldots, X_{t+h})$ for any $h \ge 0$ and measurable functions f, g is stationary and inherits the strong mixing property with the same rate as (Y_t). This immediately follows from the definitions of stationarity and strong mixing. In particular, the sequences of powers (σ_t^p) and $(|X_t|^p)$ for any positive p have mixing rates comparable to those of (σ_t) and (X_t), respectively.

3 The Covariance Structure

A first check of the dependence structure in a stationary sequence (A_t) usually focuses on the ACVF γ_A or the ACF ρ_A, see (3). Since a stochastic volatility process (X_t) is a highly non-Gaussian process its covariance structure is not very informative. In fact, as shown below, (X_t) is uncorrelated yet dependent. Insight about the nature of the dependence in stochastic volatility processes can be obtained by studying the correlation function of powers of the process and volatility process given by $(|X_t|^p)$ and (σ_t^p) for some $p > 0$, respectively.

In what follows, we focus on volatility sequences (σ_t) of the form (2) with iid random variables η_i since it has the attractive property that we get explicit representations for γ_X. Assuming $EZ = 0$, $\mathrm{var}(Z) = 1$, and $E \exp\{2|Y|\} < \infty$, direct calculation yields

$$\gamma_X(h) = \rho_X(h) = 0\,, \quad h > 0\,, \quad \mathrm{var}(X) = \mathrm{var}(\sigma)\,, \tag{6}$$

$$\mathrm{var}(\sigma) = \prod_{j=0}^{\infty} m_\eta(2\,\psi_j) - \prod_{j=0}^{\infty} m_\eta^2(\psi_j)\,, \tag{7}$$

where $m_\eta(z) = E\mathrm{e}^{\,z\eta}$ denotes the moment generating function of η. If η is centered normal with variance $\tau^2 > 0$, then we have $m_\eta(z) = \exp\{0.5\,\tau^2\,z^2\}$. Hence

$$\mathrm{var}(\sigma) = \exp\{\mathrm{var}(Y)\}(\exp\{\mathrm{var}(Y)\} - 1)\,. \tag{8}$$

We observe that (X_t) is a white noise sequence. This fact is very much in agreement with real-life return data. However, this observation is not very informative. Therefore it has become common in financial time series analysis to study the ACVFs/ACFs of the absolute values, squares and other powers of absolute return data as well. The present state of research on GARCH processes does not allow one to get explicit formulae for the ACVF of the absolute returns. In a stochastic volatility model one can exploit the independence between (σ_t) and (Z_t) in order to get explicit formulae for $\gamma_{|X|}$ at least in the model (2) with iid noise (η_i).

In what follows, we focus on this model with iid centered normal noise with variance $\tau^2 > 0$ and calculate the corresponding ACVFs and ACFs. Recall that the ACVF of (Y_t) is then given by

$$\gamma_Y(h) = \tau^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+h}\,, \quad h \geq 0\,. \tag{9}$$

Calculations similar to those leading to equations (6) and (8) yield for any $p > 0$ and $h > 0$,

$$\begin{aligned}
\gamma_{|X|^p}(h) &= (E(|Z|^p))^2\,\gamma_{\sigma^p}(h)\,,\\
\gamma_{\sigma^p}(h) &= E\mathrm{e}^{\,p\,(Y_0+Y_h)} - (E\mathrm{e}^{\,p\,Y})^2\\
&= \exp\Big\{(p\tau)^2 \sum_{i=0}^{\infty} \psi_i^2\Big\}\Big[\exp\Big\{(p\tau)^2 \sum_{i=0}^{\infty} \psi_i\,\psi_{i+h}\Big\} - 1\Big]\\
&= \mathrm{e}^{\,p^2\,\gamma_Y(0)}\Big[\mathrm{e}^{\,p^2\,\gamma_Y(h)} - 1\Big]\,,
\end{aligned}$$

where we have assumed that $E(|Z|^p) < \infty$. Since $\gamma_Y(h) \to 0$ as $h \to \infty$ a Taylor series argument yields[4]

[4] We write $f(h) \sim g(h)$ as $h \to \infty$ for two functions f and g whenever $f(h)/g(h) \to 1$ as $h \to \infty$.

$$\gamma_{\sigma^p}(h) \sim p^2 \, \mathrm{e}^{\, p^2 \, \gamma_Y(0)} \, \gamma_Y(h) \,, \quad h \to \infty \,. \tag{10}$$

Similarly,

$$\mathrm{var}(\sigma^p) = \mathrm{e}^{\, p^2 \, \gamma_Y(0)} \left(\mathrm{e}^{\, p^2 \, \gamma_Y(0)} - 1 \right),$$

$$\mathrm{var}(|X|^p) = E(|Z|^{2p}) \, \mathrm{e}^{\, p^2 \, \gamma_Y(0)} \left(\mathrm{e}^{\, p^2 \, \gamma_Y(0)} - 1 \right) + \mathrm{var}(|Z|^p) \, \mathrm{e}^{\, p^2 \, \gamma_Y(0)} \,,$$

where we have assumed that $E(|Z|^{2p}) < \infty$. Finally, we can calculate the ACFs of $(|X_t|^p)$ and (σ_t^p) for $h > 0$:

$$\rho_{\sigma^p}(h) = \frac{\mathrm{e}^{\, p^2 \, \gamma_Y(h)} - 1}{\mathrm{e}^{\, p^2 \, \gamma_Y(0)} - 1} \sim \frac{p^2}{\mathrm{e}^{\, p^2 \, \gamma_Y(0)} - 1} \, \gamma_Y(h) \,, \tag{11}$$

$$\rho_{|X|^p}(h) = \frac{(E(|Z|^p))^2}{E(|Z|^{2p}) + \mathrm{var}(|Z|^p) \, (\mathrm{e}^{\, p^2 \, \gamma_Y(0)} - 1)^{-1}} \, \rho_{\sigma^p}(h) \,. \tag{12}$$

The ACVF (9) of the linear causal Gaussian process (Y_t) in (2) may decay to zero arbitrarily fast. In particular, if (Y_t) is a causal ARMA process, the ACVF decays to zero at an exponential rate. On the other hand, if (Y_t) is a FARIMA(p, d, q) process with $d \in (0, 1)$, $\gamma_Y(h) \sim \mathrm{const}\, h^{d-1}$. In particular, the sequence (Y_t) exhibits *long-range dependence* or *long memory* in the sense that the ACVF is not absolutely summable. Otherwise, as in the ARMA case, the sequence (Y_t) is referred to as a process with *short-range dependence* or *short memory*. We refer to Brockwell and Davis (1991) and Samorodnitsky and Taqqu (1994) for extensive discussions on long memory processes, in particular on FARIMA processes and their properties. See also the more recent treatment of long memory phenomena in Doukhan et al. (2004).

We conclude from the discussion above and from formulae (10)–(12) that γ_{σ^p} inherits the asymptotic behavior of the ACVF γ_Y and, in turn, $\gamma_{|X|^p}$ inherits the asymptotic behavior of γ_{σ^p}. Since γ_Y may decay to zero at any rate we conclude that the ACVF of the processes $(|X_t|^p)$ may decay to zero at any rate as well. This allows one to model the ACVF behavior of an absolute return series in a flexible way, in contrast to the GARCH case. Indeed, under general conditions on the noise, a GARCH process (A_t) is ϕ-mixing with a geometric rate, see Mokkadem (1990), Boussama (1998), cf. Doukhan (1994). In particular, it is strongly mixing with a rate function (α_t) which decays to zero exponentially fast. Then an appeal to (4) yields for any measurable function f on $\mathbb{R}$ that $\gamma_{f(A)}(h) \to 0$ at an exponential rate as $h \to \infty$, given that $\gamma_{f(A)}$ is well defined.

4 Moments and Tails

In this section we consider some of the marginal distributional characteristics of a stochastic volatility process. It is straightforward that for any $p > 0$,

$$E(|X|^p) = E(|Z|^p)\, E\sigma^p\,,$$

and this pth moment of X is finite if and only if the pth moments of Z and σ are finite. This naturally leads to some restrictions on the moments of the noise (η_i) in model (2): the tails of η must not be too heavy, otherwise $E\sigma^p = \infty$ for all $p > 0$. This excludes in particular subexponential distribution for η for which $m_\eta(p) = \infty$ for all $p > 0$. The subexponential class includes distributions with a power law tail (such as the student and Pareto distributions) as well as moderately heavy-tailed distributions such as the Weibull distribution $P(\eta > x) = \exp\{-c\,x^\tau\}$, $x > 0$, for some $\tau \in (0, 1)$, $c > 0$, and the log-normal distributions. We refer to Embrechts et al. (1997) for an extensive treatment of subexponential distributions .

In various cases the analysis of the moments of a stochastic volatility model can be refined by a study of the asymptotic tail behavior of the distribution of X. The close relation between the moments and the tails can be seen e.g., from the fact that for any non-negative random variable A,

$$EA = \int_0^\infty P(A > x)\, dx\,. \tag{13}$$

Our particular interest focuses on non-negative random variables with power law tails of the form

$$P(A > x) = x^{-\alpha}\, L(x)\,, \quad x > 0\,, \tag{14}$$

where $\alpha > 0$ and L is a *slowly varying function* which is defined by the asymptotic relation $L(cx)/L(x) \to 1$ as $x \to \infty$, for all $c > 0$. The class of slowly varying functions includes constants, logarithms, iterated logarithms, powers of logarithms. Since for every $\delta > 0$ there exist positive constants x_0 and c_1, c_2 such that

$$c_1\, x^{-\delta} \le L(x) \le c_2\, x^\delta\,, \quad x \ge x_0\,, \tag{15}$$

the contribution of L to the tail in (14) is negligible compared to the power law $x^{-\alpha}$. The function on the right-hand side of (14) is said to be *regularly varying with index* $-\alpha$, and we will also say that the *distribution* of A is *regularly varying with index* α. It is an easy exercise to combine relations (15) and (13) to conclude that

$$E(A^{\alpha+\delta}) \begin{cases} < \infty\,, & \delta > 0\,, \\ = \infty\,, & \delta < 0\,, \end{cases}$$

whereas $E(A^\alpha)$ may be finite or infinite, depending on the slowly varying function L. For an extensive treatment of slowly varying and regularly varying functions and distributions and their properties, we refer the reader to the classical encyclopedia by Bingham et al. (1987) and to Resnick (1987).

From the definition it is evident that regularly varying distributions have heavy tails especially for small α. Therefore they are capable of capturing the probabilities of rare erratic events such as crashes, eruptions, bursts, and other phenomena which cannot be adequately described by commonly used distributions such as the normal, exponential and gamma. Examples of regularly varying distributions include the Pareto and Burr distributions which are standard models for large claims in (re)insurance applications (see Embrechts et al. (1997), Chapter 2), the ON-OFF distributions of Internet teletraffic models (see Leland et al. (1993), Willinger et al. (1995), Mikosch et al. (2002)), the one-dimensional marginals of GARCH and infinite variance stable processes (see Goldie (1991), Embrechts et al. (1997) for the tails of GARCH processes, Feller (1971) and Samorodnitsky and Taqqu (1994) for the tails of stable processes). There exists empirical evidence that the distribution of log-returns is well approximated in its left and right tails by regularly varying distributions (possibly with different tail indices on the left and on the right) such as the generalized Pareto distribution with positive shape parameter (see Embrechts et al. (1997), Chapter 6, Mikosch (2003)).

The observation that log-return data have power law tails goes back at least to the 1960s. For example, Mandelbrot (1963) suggested that infinite variance stable distributions might be appropriate models. The latter class of distributions is regularly varying with index $\alpha < 2$, see Feller (1971) and Samorodnitsky and Taqqu (1994). Since Mandelbrot's contributions were not based on rigorous statistical analysis there has been an ongoing discussion about the value α and whether power law tails make sense for financial data at all. A detailed statistical analysis of the tail index α of return data depends on conditions such as strict stationarity which is unlikely to be satisfied for large samples, whereas the estimation of α requires large samples (sizes of 1000 observations and more are desirable, see Embrechts et al. (1997), Chapters 6 and 7). Since changes in the distribution of returns are likely in large samples it is a rather difficult task to decide a value of α that is appropriate for the entire segment of a long series. Nevertheless, there is a strong belief by many researchers that return data have power law tails. Using extreme value statistics for estimating the tail index α, one often finds α to be well below 5 or 6 yet greater than 2.

Since returns may have heavy-tailed distributions it is natural to ask for conditions on the stochastic volatility process which ensure the existence of power law tails in the marginal distribution. Recall that $X_t \stackrel{d}{=} X = \sigma Z$, where $\sigma_t \stackrel{d}{=} \sigma$ and $Z_t \stackrel{d}{=} Z$ are independent random variables. In this context it is useful to recall an elementary result of Breiman (1965). Let A, B be non-negative random variables such that A is regularly varying with index α and $E(B^{\alpha+\delta}) < \infty$ for some $\delta > 0$. Then

$$P(A\,B > x) \sim E(B^\alpha)\,P(A > x)\,, \quad x \to \infty\,.$$

In particular, AB is regularly varying with index α. Thus the product inherits the heavier tail of the two factors.[5]

An immediate consequence is that

$$P(X > x) = P(Z_+\,\sigma > x) \sim E(\sigma^\alpha)\,P(Z_+ > x)\,, \quad x \to \infty\,, \tag{16}$$

provided $Z_+ = \max(0, Z)$ has a regularly varying distribution with index $\alpha > 0$ and $E(\sigma^{\alpha+\delta}) < \infty$ for some $\delta > 0$. The latter condition is satisfied in model (2) with iid normal noise. Then σ is lognormal, hence all moments $E(\sigma^p)$, $p > 0$, are finite. Analogously,

$$P(X \le -x) = P(Z_-\,\sigma \ge x) \sim E(\sigma^\alpha)\,P(Z_- > x)\,, \quad x \to \infty\,,$$

provided $Z_- = \max(0, -Z)$ has a regularly varying distribution with index α and $E(\sigma^{\alpha+\delta}) < \infty$ for some $\delta > 0$. The case of σ and Z with heavy tails of the same order of magnitude is rather involved. As a matter of fact, if both σ and Z_+ are regularly varying with index $\alpha > 0$, then X_+ is regularly varying with the same index but the form of the slowly varying function L in the tail is in general not known, see Embrechts and Goldie (1980).

Breiman's result (16) tells us that a power law tail for X may result from a heavy tail of the volatility σ or of the noise Z. Since we observe neither σ nor Z we can only judge about their distributional tail on the basis of a model such as the stochastic volatility model or the GARCH process. In the GARCH case power law tails of $P(X_t > x)$ are more the rule than the exception:[6] even for light-tailed Z (such as Gaussian noise) the volatility σ will typically have power law tails.

The tail behavior of the marginal distribution of a stationary sequence (X_t) is essential for its extremal behavior. In particular, power law behavior for the tail $P(X > x)$ often results in the fact that scaled maxima $\max_{t=1,\dots,n} X_t$ converge in distribution to a Fréchet distributed random variable. We study the convergence of the extremes of a stochastic volatility process in Davis and Mikosch (2008b). There we consider the case of regularly varying X, but also some light-tailed X and the corresponding extreme value theory.

5 Asymptotic Theory for the Sample ACVF and ACF

In this section we briefly study the asymptotic behavior of the sample mean, and the sample ACVFs of the stochastic volatility process (X_t), its absolute

[5] Of course, $E(B^{\alpha+\delta}) < \infty$ for some $\delta > 0$ and regular variation of A with index α imply that $P(B > x) = o(P(A > x))$ as $x \to \infty$.

[6] See the article about the extremes of a GARCH process in Davis and Mikosch (2008a).

values ($|X_t|$) and its squares. Recall that the *sample ACVF* and the *sample ACF* of a stationary sequence (A_t) are given by

$$\widehat{\gamma}_A(h) = \frac{1}{n}\sum_{t=1}^{n-h}(A_t - \overline{A}_n)\,(A_{t+h} - \overline{A}_n)\,, \quad \widehat{\rho}_A(h) = \frac{\widehat{\gamma}_A(h)}{\gamma_A(0)}\,, \quad 0 \le h < n\,,$$

respectively, where $\overline{A}_n = n^{-1}\sum_{t=1}^{n} A_t$ denotes the mean of the sample $A_1, \ldots, A_n$. If (σ_t), hence (X_t), is stationary ergodic, then the ergodic theorem (cf. Krengel (1985)) implies that the sample ACVFs at a fixed lag h, $\widehat{\gamma}_\sigma(h), \widehat{\gamma}_X(h), \widehat{\gamma}_{|X|^i}(h)$, $i = 1, 2$, converge a.s. to their corresponding deterministic counterparts $\gamma_\sigma(h)$, $\gamma_X(h)$, $\gamma_{|X|^i}(h)$, $i = 1, 2$, provided that the limiting covariances exist and are finite. The corresponding sample ACFs at a fixed lag h will then converge a.s. to their deterministic counterparts as well.

Central limit theory for functionals of a stochastic volatility process (X_t) is often easily established. In what follows, we give some examples which are not exhaustive but illustrative of the techniques that are involved. Assume that the log-volatility process (Y_t) is given by the representation (2) for an iid sequence (η_i) and that $\mathrm{var}(\sigma) < \infty$, $EZ = 0$ and $\mathrm{var}(Z) = 1$. We introduce the filtration $\mathcal{G}_t = \sigma(Z_s\,, \eta_s\,, s \le t)$. Then (X_t) is adapted to $(\mathcal{G}_t)$, $\mathrm{var}(X) < \infty$ and (recall that $\psi_0 = 1$)

$$E(X_t \mid \mathcal{G}_{t-1}) = \mathrm{e}^{\sum_{i=1}^{\infty}\psi_i\,\eta_{t-i}}\,E(Z_t\,\mathrm{e}^{\,\eta_t}) = 0 \quad \text{a.s.}$$

Hence (X_t) constitutes a finite variance martingale difference sequence and therefore the central limit theorem for stationary ergodic martingale sequences applies (see Billingsley (1968)):

$$\sqrt{n}\,\overline{X}_n \xrightarrow{d} N(0, E(\sigma^2))\,.$$

Similarly, for $h > 0$, $(X_t X_{t+h})$ is adapted to the filtration $(\mathcal{G}_{t+h})$, and if in addition $E(\sigma^4) < \infty$ we have

$$\mathrm{var}(X_t X_{t+h}) = E(\sigma_0^2\sigma_h^2) < \infty$$

and

$$E(X_t\,X_{t+h} \mid \mathcal{G}_{t+h-1}) = X_t\,E(X_{t+h} \mid \mathcal{G}_{t+h-1}) = 0 \quad \text{a.s.}$$

Therefore $(X_t X_{t+h})$ is a mean zero finite variance stationary ergodic martingale difference sequence. The ergodic theorem and the central limit theorem for stationary martingale differences yield

$$\sqrt{n}\,\widehat{\gamma}_X(h) = \sqrt{n}\sum_{t=1}^{n} X_t\,X_{t+h} - \sqrt{n}(\overline{X}_n)^2 + o_P(1)$$

$$= \sqrt{n} \sum_{t=1}^{n} X_t \, X_{t+h} + o_P(1)$$

$$\xrightarrow{d} N(0, E(\sigma_0^2 \sigma_h^2)) \, .$$

Central limit theory can be derived for the sample means of the absolute values and squares of (X_t) as well as for $\widehat{\gamma}_{|X|^i}$, $i = 1, 2$, under additional strong mixing conditions. In Section 2.2 we have learned that (X_t) inherits strong mixing with a certain rate function (α_t) from the log-volatility sequence (σ_t). Given that the rate condition

$$\sum_{t=1}^{\infty} \alpha_t^{\delta/(2+\delta)} < \infty \tag{17}$$

holds for some $\delta > 0$, one can apply a classical central limit theorem, see Ibragimov and Linnik (1971). Condition (17) is satisfied if $\alpha_t \to 0$ at an exponential rate. It is satisfied e.g., for Gaussian ARMA log-volatility processes (Y_t) in (2). The central limit theorem applies to any strongly mixing sequence (A_t) with rate function (α_t) satisfying the conditions (17) and $E(|A|^{2+\epsilon}) < \infty$ for some $\epsilon > 0$. In particular, it is applicable to $A_t = \sigma_t$ and $A_t = |X_t|^p$ for any $p > 0$, but also for $A_t = |X_t X_{t+h}|^p$ for any $p > 0$. We omit further details.

It is also possible to derive limit theory with non-Gaussian limits for the ACVFs/ACFs of stochastic volatility processes (X_t), their absolute values and squares when standard moment conditions such as $\mathrm{var}(X) < \infty$ fail. Davis and Mikosch (2001a) (see also Davis and Mikosch (2001b)) prove for regularly varying Z with index $\alpha \in (0, 2)$ and a Gaussian log-volatility process (Y_t) in (2) that the scaled sample ACVF $\widehat{\gamma}_X(h)$ at the fixed lag $h \geq 0$ converges in distribution to an infinite variance α-stable limit (see Samorodnitsky and Taqqu (1994) for a discussion of stable distributions) at a rate which depends on the tail of Z. Notice that in this case, X is regularly varying with index α and therefore $\mathrm{var}(X) = \infty$, see Section 4. In particular, the notions of ACVF/ACF are not defined. However, the sample ACF at a fixed lag h, $\widehat{\rho}_X(h)$, converges to zero even when the ACF is not defined. The rate at which this convergence happens is of the order $n^{1/\alpha}$, hence it is much faster than the common $\sqrt{n}$-rates for Gaussian central limit theory. Analogous results apply to the sample ACFs of the absolute values $|X_t|$ and the squares X_t^2. In the case of the squares one has to alter the condition of regular variation: Z must be regularly varying with index $\alpha \in (0, 4)$. Since, on the one hand, it has become common to study the sample ACFs of squared returns and, on the other hand, return series may have infinite moments of low order, the limit theory for the sample ACVFs/ACFs can be quite important in situations when one lacks sufficiently high moments.

We mention in passing that the limit theory for the sample ACVFs/ACFs of a stochastic volatility process, its absolute values and squares very much

parallels the corresponding theory for an iid sequence, also in the infinite variance situation. Moreover, the limit theory for a heavy-tailed GARCH process (X_t) is of a completely different nature, see Davis and Mikosch (1998), Mikosch and Stărică (2000) and Basrak et al. (2002); cf. Davis and Mikosch (2001b) and Mikosch (2003) for overviews. In particular, if the marginal distribution of a GARCH process is regularly varying with index $\alpha \in (2, 4)$ the sample ACF $\widehat{\rho}_X$ converges to 0 at a rate much slower than $\sqrt{n}$ and if $\alpha \in (0, 2)$ the sample ACF has a non-degenerate limit distribution without any normalization. The latter property would lead to completely different graphs for the sample ACFs for disjoint time intervals. This is another property which highlights a crucial difference between the stochastic volatility and GARCH models for returns.

References

Basrak, B., Davis, R.A. and Mikosch, T. (2002): Regular variation of GARCH processes. *Stoch. Proc. Appl.* **99**, 95–116.

Billingsley, P. (1968): *Convergence of Probability Measures.* Wiley, New York.

Billingsley, P. (1995): *Probability and Measure.* 3rd edition. Wiley, New York.

Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987): *Regular Variation.* Cambridge University Press.

Bougerol, P. and Picard, N. (1992): Stationarity of GARCH processes and of some nonnegative time series. *J. Econometrics* **52**, 115–127.

Boussama, F. (1998): *Ergodicité, mélange et estimation dans le modelès GARCH.* PhD Thesis, Université 7 Paris.

Bradley, R. (2005): Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* **2**, 107–144.

Breiman, L. (1965): On some limit theorems similar to the arc-sine law. *Theory Probab. Appl.* **10**, 323–331.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods* (2nd edition). Springer, Berlin, Heidelberg, New York.

Davis, R.A. and Mikosch, T. (1998): Limit theory for the sample ACF of stationary process with heavy tails with applications to ARCH. *Ann. Statist.* **26**, 2049–2080.

Davis, R.A. and Mikosch, T. (2001a): Point process convergence of stochastic volatility processes with application to sample autocorrelations. *J. Appl. Probab. Special Volume: A Festschrift for David Vere-Jones* **38A**, 93–104.

Davis, R.A. and Mikosch, T. (2001b): The sample autocorrelations of financial time series models. *In: Fitzgerald, W.J., Smith, R.L., Walden, A.T. and Young, P.C. (Eds.): Nonlinear and Nonstationary Signal Processing*, 247–274. Cambridge University Press.

Davis, R.A. and Mikosch, T. (2008a): Extreme value theory for GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 186–200. Springer, New York.

Davis, R.A. and Mikosch, T. (2008b): Extreme value theory for stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 355–364. Springer, New York.

Doukhan, P. (1994): *Mixing. Properties and Examples. Lecture Notes in Statistics* **85**, Springer, New York.

Doukhan, P., Oppenheim, G. and Taqqu, M.S. (Eds.) (2004): *Long Range Dependence* Birkhäuser, Boston.

Eberlein, E. and Taqqu, M.S. (Eds.) (1986): *Dependence in Probability and Statistics.* Birkhäuser, Boston.

Embrechts, P. and Goldie, C.M. (1980): On closure and factorization theorems for subexponential and related distributions. *J. Austral. Math. Soc. Ser. A* **29**, 243–256.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997): *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

Fan, J. and Yao, Q. (2003): *Nonlinear Time Series.* Springer, Berlin, Heidelberg, New York.

Feller, W. (1971): *An Introduction to Probability Theory and Its Applications II.* Wiley, New York.

Goldie, C.M. (1991): Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.* **1**, 126–166.

Ibragimov, I.A. and Linnik, Yu.V. (1971): *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

Krengel, U. (1985): *Ergodic Theorems.* De Gruyter, Berlin.

Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson, D.V. (1993): On the self-similar nature of Ethernet traffic. *ACM/SIGCOMM Computer Communications Review*, 183–193.

Lindner, A.M. (2008): Stationarity, Mixing, Distributional Properties and Moments of GARCH(p, q)–Processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 43–69. Springer, New York.

Mandelbrot, B. (1963): The variation of certain speculative prices. *J. Busin. Univ. Chicago* **36**, 394–419.

Mikosch, T. (2003): Modelling dependence and tails of financial time series. *In: Finkenstädt, B. and Rootzén, H. (Eds.): Extreme Values in Finance, Telecommunications and the Environment*, 185–286. Chapman and Hall.

Mikosch, T., Resnick, S., Rootzén, H. and Stegeman, A. (2002): Is network traffic approximated by stable Lévy motion or fractional Brownian motion? *Ann. Appl. Probab.* **12**, 23–68.

Mikosch, T. and Stărică, C. (2000): Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Ann. Stat.* **28**, 1427–1451.

Mokkadem, A. (1990): Propriétés de mélange des processus autoregréssifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist.* **26**, 219–260.

Nelson, D.B. (1990): Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory* **6**, 318–334.

Petrov, V.V. (1995): *Limit Theorems of Probability Theory.* Oxford University Press, Oxford.

Pham, T.D. and Tran, L.T. (1985): Some mixing properties of time series models. *Stoch. Proc. Appl.* **19**, 279–303.

Resnick, S.I. (1987): *Extreme Values, Regular Variation, and Point Processes.* Springer, New York.

Samorodnitsky, G. and Taqqu, M.S. (1994): *Stable Non–Gaussian Random Processes. Stochastic Models with Infinite Variance.* Chapman and Hall, London.

Willinger, W., Taqqu, M.S., Sherman, R. and Wilson, D. (1995): Self-similarity through high variability: statistical analysis of ethernet lan traffic at the source level. *Proceedings of the ACM/SIGCOMM'95, Cambridge, MA. Computer Communications Review* **25**, 100-113.

Moment–Based Estimation of Stochastic Volatility Models

Eric Renault*

Abstract This chapter reviews the possible uses of the Generalized Method of Moments (GMM) to estimate Stochastic Volatility (SV) models. A primary attraction of the method of moments technique is that it is well suited for identifying and estimating volatility models without a complete parametric specification of the probability distributions. Moreover, simulation-based methods of moments are able to exploit a variety of moments, while avoiding limitations due to a lack of closed form expressions. The chapter first highlights the suitability of GMM for popular regression models of volatility forecasting. Then, it reviews the implications of the SV model specification in terms of higher order moments: skewness, kurtosis, variance of the variance, leverage and feedback effects. The chapter examines the ability of a continuous time version of SV models to accommodate data from other sources like option prices or high frequency data on returns and transactions dates. Simulation-based methods are particularly useful for studying continuous time models due to the frequent lack of closed form expressions for their discrete time dynamics. These simulation-based methods of moments are presented within the unifying framework of indirect inference with a special emphasis on misspecification. Likely misspecification of the parametric model used for simulation requires a parsimonious and well-focused choice of the moments to match.

Eric Renault
University of North Carolina, Department of Economics, Gardner Hall, Chapel Hill, NC 27599 e-mail: renault@email.unc.edu

* We would like to thank Mike Aguilar, Eric Ghysels, and a referee for insightful comments. Most especially, we would like to thank Torben Andersen for many useful suggestions.

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_12,

1 Introduction

Estimators based on the Method of Moments (MM) or the Generalized Method of Moments (GMM) have been widely applied since the early days of the Stochastic Volatility (SV) literature. There are at least two explanations for the popularity of these approaches. First, moments of financial time series have always been of primary interest as such moments are associated not only with volatility forecasting but also important aspects of the return distribution such as heavy tails and return-volatility asymmetries, see e.g., Rosenberg (1972) and Black (1976). Second, besides modeling issues, MM approaches are popular for their simplicity as the exact likelihood function is difficult to evaluate in a context of parametric volatility models within the class of hidden Markov models.

The simplicity argument often raises the criticism of lack of statistical efficiency (see e.g. Shephard (2005), p. 13). However, in the context of highly nonlinear dynamic models with fat tails and latent variables, asymptotic efficiency of the maximum likelihood (ML) estimator is not always warranted. Moreover, when efficient estimation is a well-defined target, there are at least two different arguments for the defense of moment-based estimation.

Tractable analytic expressions for moments are no longer an issue since simulations may provide accurate Monte Carlo counterparts to any moment of interest. The benefits of a general Simulated Method of Moments (SMM) approach have been widely documented in the last decade by the literature on Indirect Inference (I.I) and Efficient Method of Moments (EMM) initiated by Smith (1993) and later developed by Gouriéroux, Monfort and Renault (1993) (GMR) and Gallant and Tauchen (1996) (GT). As noted by Shephard (2005), "throughout their development of this rather general fully parametric method, both GMR and GT had very much in mind the task of performing reasonably efficient inference on SV models". The key point is that the efficiency of ML is reached by MM whenever the chosen moments are rich enough to span the likelihood score vector (see Carrasco and Florens (2002) and Carrasco, Chernov, Florens and Ghysels (2007) for a recent reappraisal of this classical result). The value added by simulation based approaches is that we may exploit a variety of moments without limitations due to a lack of closed-form expressions.

A second reason why alleged statistical inefficiency of MM may not be a sufficient motivation to abandon it is closely related to the motivation of SV modeling. It is natural to identify and estimate volatility models through their ability to reproduce the moments of interest which they have been designed to capture. Of course, GMM implementations should select the set of moments intelligently to improve estimation accuracy. However, the moments of financial interest should remain at the core of the procedure, both for meaningful statistical identification as well as robustness to misspecification. In fact, multivariate SV models may be spuriously identified by assumptions about probability distributions that have little to do with the moment struc-

ture these models are supposed to focus on. Generally, financial theory has little to say about moments of orders beyond two, three or four. Moreover, financial loss functions are typically less concerned with efficient inference than with robust estimation of some moments of interest related for instance to risk premiums, Value at Risk, and expected shortfall. In this respect, by performing allegedly efficient estimation in a fully parametric model which inevitably has been specified rather arbitrarily, one runs the risk of contaminating the estimation of the parameters of interest through likely misspecification of some features of the model. This is a reason why semi-parametric identification and inference, as provided by MM, might be an attractive option for many purposes in the context of SV models.

Note that the two above arguments may seem to be at odds. Point one requires a specific data generation process for simulation whereas the second point focuses on semi-parametric procedures. The two points must be essentially understood as applying to different circumstances. However, it is to some extent possible to devise a simulation-based MM while maintaining the semi-parametric philosophy of GMM (see e.g. Dridi, Guay and Renault (2007)). Consequently, there are several reasons to view moment based estimation as a valuable tool and not simply a naive method employed because efficient inference is too complicated. We organize the review of relevant MM estimation methods according to their objectives.

Section 2 focuses on parameters of interest for volatility forecasting. As noted in the pioneering work of Rosenberg (1972), regression models of variance fluctuations are the right tool for that. Section 3 briefly reviews the implication of SV model specification in terms of higher order moments: skewness, kurtosis, variance of the variance, leverage and feedback effects. Section 4 is devoted to continuous time stochastic volatility models. These models often impose restrictions on the higher order return moments and facilitate the development of tools for analyzing data from other sources like option prices or high frequency data on returns and transaction dates. Section 5 is devoted to simulation-based methods of moments while some concluding remarks are assembled in Section 6.

The main limitation of this chapter is that we focus on endogenous modeling of volatility. With the notable exception of random durations between trades or quotes, we never consider observed explanatory variables for the volatility of returns other than the past realizations of returns themselves. Hence, we will not cover "economic' models of volatility such as in Schwert (1989) and more recently Engle, Ghysels and Sohn (2006). Moreover, for a comprehensive treatment of GMM estimation and inference in time series models, we refer the reader to Hall (2005). This elegant book provides the methodological details not covered here and includes an example of stochastic volatility models.

2 The Use of a Regression Model to Analyze Fluctuations in Variance

The title of this section is borrowed from Rosenberg (1972), a seminal albeit unpublished paper, recently reprinted in Shephard (2005). Shephard (2005) points out that Rosenberg (1972) is "by far the closest precursor of the ARCH class of models". We show below that what Rosenberg labeled the "regression model" actually encompasses most of the nowadays popular SV models as well as the GARCH model as a limiting case.

2.1 The linear regression model for conditional variance

The key contribution of Rosenberg (1972) is to be the first to realize that fat tails observed in asset prices changes $z_{t+1} = \log(\frac{P_{t+1}}{P_t})$ can be explained by a decomposition:

$$z_{t+1} = m_t + \sigma_t \varepsilon_{t+1} \tag{1}$$

where in his words (but different notation) " the ε_t are serially independent random variables with identical distribution function $F(\cdot)$ having mean equal to zero, variance equal to one, and kurtosis equal to κ. The variables σ_t, which are the variances of price changes, obey a stochastic process that can be forecasted. The ε_{t+1} are contemporaneously independent of the σ_t ".

Then, in strikingly modern terms, he notes that "if the variance (...) was an observable variable, it would be possible to regress it directly upon explanatory variables (...). In reality, the variance is not observed, but the observed squared price change is a realization of the underlying distribution with expected value equal to the variance and, accordingly, provides a means to carry out the regression". Starting from a vector $x_t = (x_{kt})_{1 \leq k \leq K}$ of K explanatory (or predetermined) variables, he writes down an "additive model":

$$\sigma_t^2 = x_t' b + u_t \tag{2}$$

where the disturbances u_t are assumed to be serially independent with mean zero and variance σ_u^2. He also mentions explicitly that "in the absence of operational measurements of the exogenous factors influencing this variance", it is possible "to employ a moving average of realized squared price changes as an estimate of the prevailing variance", leading to the model:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i y_{t+1-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + u_t \tag{3}$$

where

$$y_{t+1} = z_{t+1} - m_t = \sigma_t \varepsilon_{t+1} \tag{4}$$

is the innovation of the return process and ω denotes an unknown intercept. In particular in the case where u_t is exactly zero, he produces the GARCH(p,q). He explicitly mentions this degenerate case by noting that it gives an upper bound to the kurtosis of the standardized return innovation ε_{t+1} for a given value of the kurtosis of the return innovation y_{t+1}. Expresssed differently, he anticipates the key comparison by Kim, Shephard and Chib (1998) between a genuine SV model (non-zero u_t) with Gaussian innovations and an heavy tailed version of the GARCH model.

Of course, as noted by Shephard (2005), what was missing in Rosenberg (1972) and was the main insight of Engle (1982), is that "the degenerate GARCH case is key as it produces a one-step ahead conditional model for returns given past data, which (...) immediately yields a likelihood function". By contrast, when the error term u_t in the regression equation (3) is non zero, this equation only provides what Drost and Nijman (1993) have called a weak GARCH(p,q) : the linear projection σ_t^{*2} of σ_t^2 (or equivalently, as put in Drost and Nijman (1993), of y_{t+1}^2) on past $y_{t+1-i}^2, i \geq 1$, is driven by a dynamic akin to the GARCH(p,q) model:

$$\sigma_t^{*2} = \omega + \sum_{i=1}^{p} \alpha_i y_{t+1-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^{*2} \tag{5}$$

Following Francq and Zakoïan (2000), we associate weak GARCH with a notion of volatility dynamics based on linear projections on past squared returns only, while Drost and Nijman (1993), for the sake of temporal aggregation, also consider linear projections on past returns themselves. Irrespective of this definitional choice, the above linear weakening of the GARCH property destroys its main advantage since the (quasi) maximum likelihood approach does not provide a consistent estimator anymore (see Meddahi and Renault (2004) for compelling Monte Carlo evidence). The only way to generate a consistent estimator from the weak GARCH model (5) is to note that it provides a linear ARMA$[\max(p,q),q]$ representation for squared returns since with $\zeta_{t+1} = y_{t+1}^2 - \sigma_t^{*2}$, we have:

$$y_{t+1}^2 = \omega + \sum_{i=1}^{\max(p,q)} \gamma_i y_{t+1-i}^2 - \sum_{j=1}^{q} \beta_j \zeta_{t+1-j} + \zeta_{t+1} + u_t$$

where, for the purpose of characterizing the volatility persistence, the parameters of interest are now the γ_is. In particular, $\gamma_i = \alpha_i + \beta_i$ for $i = 1, .., \min(p,q)$. Note that in general the y_ts may be return innovations rather than returns themselves but, by a common abuse of language, we call them returns from now on. As noted by Rosenberg (1972), the case where all the β_js are assumed to be zero (the weak ARCH(p) case) allows a simple GLS estimation while in the general case (non-zero β_js) there is a classical errors-in-variables problem: least squares estimation of the regression parameters

γ_is in the regression equation (8) may not be consistent due to contemporaneous correlation between the y^2_{t+1-i} and the "errors" ζ_{t+1-j}. The right way to address this estimation issue is, as shown by Francq and Zakoïan (2000), to use instead the minimization of the sum of the squared linear innovations of the ARMA process (y^2_t) as they can be computed from ζ_t and u_t . One potential drawback of any MM estimation based on the regression model (3), including the least squares (LS) approach of Francq and Zakoïan (2000), is that a standard GMM theory can be applied only if asset returns have a finite fourth moment. In fact, Francq and Zakoïan (2000) require finite fourth moment for consistency and finite eight moment for asymptotic normality of the LS estimator.

In contrast, quasi-maximum GARCH likelihood theory implies both consistency and asymptotic normality, as long as the standardized innovation ε_t has a finite fourth moment (see e.g. Francq and Zakoïan (2004)). The existence of a finite fourth moment for returns themselves is a much more restrictive condition. For instance, when the error process (u_t) in (3) itself has finite forth moment ($u_t \equiv 0$ in the GARCH case), the required condition can be expressed as an inequality restriction about the coefficients $(\alpha_i)_{1\leq i\leq p}$ and $(\beta_j)_{1\leq j\leq q}$ which coincides with the standard one for a $GARCH(p,q)$ model. In the simplest case of a conditionally Gaussian return process with $p = q = 1$, this condition (as established for GARCH$(1,1)$ by Bollerslev (1986)) is:

$$(\alpha+\beta)^2 + 2\alpha^2 < 1 \ .$$

2.2 The SR–SARV(p) model

Drost and Nijman (1993) introduce the weak GARCH concept to exploit the temporal aggregation properties of linear ARMA models. However, a drawback is that much of the useful structure of the volatility model is lost when it comes to parameter estimation. The purpose of this section is to show that a general regression model like (3) provides conditional moment restrictions much more powerful for inference than the linear ARMA structure of squared returns alone. Moreover, these conditional moment restrictions characterize a class of stochastic volatility processes, labeled SR–SARV(p), which share the advantage of weak GARCH in terms of temporal aggregation.

To develop the argument, let η_{t+1} denote the forecast error when y^2_{t+1} is forecasted by σ^2_t:

$$y^2_{t+1} = \sigma^2_t + \eta_{t+1} \tag{6}$$

$$E[\eta_{t+1}\,|y_{t-h}, h \geq 0] = 0 \tag{7}$$

Note that, in general, σ^2_t denotes the conditional expectation of y^2_{t+1} given an information set strictly larger than the one generated by values of observed

past returns y_{t-h}, $h \geq 0$, precisely because $u_t \neq 0$. However, the conditional moment restriction (7) can always be deduced by the law of iterated expectations. For the sake of inference from observed past returns, let us rewrite (3) in terms of squared returns and their forecast errors. We get the regression equation:

$$y_{t+1}^2 = \omega + \sum_{i=1}^{\max(p,q)} \gamma_i y_{t+1-i}^2 - \sum_{j=1}^{q} \beta_j \eta_{t+1-j} + \eta_{t+1} + u_t \tag{8}$$

The modern way to address the aforementioned errors-in-variables problem stressed by Rosenberg (1972)) is to deduce from (8) the following conditional moment restrictions that identify both the volatility persistence parameters γ_is and the unconditional variance $\omega[1 - \sum_{i=1}^{\max(p,q)} \gamma_i]^{-1}$:

$$E[y_{t+1}^2 - \omega - \sum_{i=1}^{\max(p,q)} \gamma_i y_{t+1-i}^2 | (y_{t-j}, j \geq q)] = 0 \tag{9}$$

The restrictions (9) belong to the class of "Multiperiod Conditional Moment Restrictions" studied by Hansen, Heaton and Ogaki (1988). Basically, they characterize the ARMA$[\max(p,q), q]$ structure of squared returns. Without heteroskedasticity in squared returns, the efficiency bound for GMM estimation of the autoregressive parameters γ_i would coincide with the asymptotic covariance matrix of the gaussian maximum likelihood. Unfortunately, the squared returns are typically highly heteroskedastic and in turn, the MA error in equation (9), $[u_t - \sum_{j=1}^{q} \beta_j \eta_{t+1-j} + \eta_{t+1}]$, will be severely heteroskedastic, rendering the characterization of the efficiency bound (see Hansen, Heaton and Ogaki (1988)) more complicated. The moment restrictions (9) represent all what we know, which is more than given by the weak-GARCH structure but less than assumed via the standard GARCH specification (which basically corresponds to $u_t \equiv 0$). Consequently, it is natural to introduce a class of stochastic volatility processes in between weak GARCH and standard GARCH, as propsed by Meddahi and Renault (2004):

Definition 1 A stationary martingale difference sequence (y_t) is a SR–SARV(p) if there exists $(p+1)$ real numbers $\omega, \gamma_1, ..., \gamma_p$, such that all the roots of the complex polynomial $[1 - \sum_{i=1}^{p} \gamma_i z^i]$ are outside the unit circle and:

$$E[y_{t+1}^2 - \omega - \sum_{i=1}^{p} \gamma_i y_{t+1-i}^2 | (y_{t-j}, j \geq p)] = 0$$

Note that the definition requires the martingale difference property only with respect to the natural filtration based on past values of (y_t). This does not prevent us from interpreting the definition with respect to a wider conditioning set J_t. Andersen (1994) considered the general class of SARV models where a function of the volatility process is a polynomial of an AR(1)

process. In the case of SR–SARV(1), it is easy to see that the volatility process σ_t is actually the square root of an AR(1) process (hence the acronym SR–SARV(1) proposed by Andersen(1994)) defined by the optimal forecast $E[y_{t+1}^2|J_t]$ of y_{t+1}^2. More generally, Meddahi and Renault (2004) show that the $SR-SARV(p)$ property can be characterized by the fact that $\sigma_t^2 = E[y_{t+1}^2|J_t]$ is a linear combination of the p components of an invertible VAR(1) process whose innovation process is a martingale difference sequence. In turn, this implies that the conditional variance process σ_t^2 is ARMA$(p, p-1)$ (σ_t^2 is AR(1) in the case of SR–SARV(1)) but provides much richer information, via the above moment restrictions than the weak linear ARMA structure for the task of MM inference.

In summary, the weak GARCH of Drost and Nijman (1993) and the SR-SARV of Meddahi and Renault (2004) are two alternative ways to relax the GARCH definition, in order to restore robustness with respect to temporal aggregation. While the former only maintains the GARCH structure for linear projections, the latter can be interpreted as maintaining it, up to some exogenous shock which makes volatility stochastic. More precisely, let us assume that:

$$\sigma_t^2 = h_t + k_t \tag{10}$$

where h_t is driven by a GARCH(p, q) equation:

$$h_t = \omega + \sum_{i=1}^{p} \alpha_i y_{t+1-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \tag{11}$$

and, for sake of expositional simplicity, (k_t) is a non-negative i.i.d. sequence independent from the process (h_t). Then, we again obtain Rosenberg's regression model (3) with:

$$u_t = k_t - \sum_{j=1}^{q} \beta_j k_{t-j} - (1 - \sum_{j=1}^{q} \beta_j) E(k_t)$$

Since the regression model (3) implies the conditional moment restriction (9) which in turn corresponds to the SR–SARV(max(p, q)) property, a possible interpretation of the latter is a GARCH plus shock model (10) and (11). This interpretation is used (in the particular case $p = q = 1$) in Franses, Van der Leij and Paap (2008) to test for GARCH$(1, 1)$ within the more general stochastic volatility model by testing that the variance of the shock k_t is zero. In particular, they find that fitting a GARCH$(1, 1)$ when it is rejected tends to inflate the estimates controlling the fatness of the tail for the standardized innovations, in accordance with Rosenberg's comment on the limiting case $u_t \equiv 0$. While Franses, Van der Leij and Paap (2008) run this test in a parametric likelihood framework, Aguilar and Renault (2008) perform it with a MM approach based on (9) as well as higher order conditional moment restrictions to capture the excess kurtosis induced by a non-zero u_t. These

higher order moments are described in Section 3 below. Although beyond the scope of this chapter, the likelihood-based test of EGARCH versus stochastic volatility proposed by Kobayashi and Shi (2005) is also best interpreted in the context of Exponential SARV which we study next.

2.3 The Exponential SARV model

As stressed by Engle (1995) and highlighted in the previous section, estimating volatility models by time-series techniques via ARMA models for squared returns is generally very inefficient since these models feature "innovations sequences which not only have time-varying variances but have bounded support which differs over time". This may motivate a preliminary log-transformation of the conditional variance, as also proposed by Rosenberg (1972), before fitting an ARMA model. For instance, while σ_t^2 was AR(1) in the case of the SR–SARV(1) model studied above, Taylor (1986) put forth the so-called log-normal $SARV(1)$ model defined as:

$$\begin{aligned} y_{t+1} &= \sigma_t \varepsilon_{t+1} \\ \log \sigma_t^2 &= \omega + \phi \log \sigma_{t-1}^2 + u_t \end{aligned} \tag{12}$$

where the innovation u_t of the volatility equation is assumed to be a Gaussian white noise process with variance σ_u^2. While the assumption of Gaussianity of u_t justifies the terminology of log-normal SARV model, we can more generally interpret this model semiparametrically and, following Andersen (1994), denote it Exponential-SARV, since now the volatility process appears as the exponential of an AR(1) process. Although it is often assumed that a distributional assumption is necessary for estimation of the SARV model (12), Francq and Zakoïan (2006) point out this is not necessary since:

1. As already noted by Ruiz (1994), a simple transformation of (12) as:

$$\log y_{t+1}^2 = \log \sigma_t^2 + \log \varepsilon_{t+1}^2$$

   allows to see $\log y_{t+1}^2$ as an AR(1) plus noise that is an $ARMA(1,1)$. This ARMA(1, 1) representation allows Francq and Zakoïan to get a consistent LS estimator of the parameters of interest up to an arbitrary value for the expectation of $\log \varepsilon_{t+1}^2$. This variable is clearly defined only up to some additive constant because ε_{t+1} is defined up to a scaling factor.
2. The main result of Francq and Zakoïan (2006) is that an $ARMA(m,m)$ model is valid for any power $[\log y_{t+1}^2]^m$, $m = 1, 2, \ldots$. This remark allows them to devise more accurate weighted LS estimators in the spirit of overidentified GMM.

Of course, there is no such thing as a free lunch. Deciding to write a regression equation in terms of log-variance instead of variance comes at a

price which is the loss of a versatile semi-parametric link between the linear correlation structure of squared returns (rather than log-square returns) and moments of their unobserved conditional variance. It is however worth knowing that, when considering the necessity of a distributional assumption, the log-normality of conditional variance appears to be a rather realistic one according to Andersen et al. (2001, 2003).

The simple formulas of moment of a log-normal distribution actually allow to easily compute, when u_t in (12) is log-normal, all the autocorrelations $\rho_{y^2}(j)$ at any positive lag j between squared returns y_t^2 and y_{t-j}^2 and to show that $\rho_{y^2}(j)$ is proportional to $[\exp(\xi\phi^j) - 1]$ where $\xi = Var(\log \sigma_t^2) = \frac{\sigma_u^2}{1-\phi^2}$. Taylor (1986), page 75, argues that this autocorrelation function is very similar to something like $C\phi^j$ with $0 < C < 1$, that is the autocorrelation function of an $ARMA(1,1)$ process with positive roots for both the AR and MA polynomials. In other words, the autocorrelogram of the squared return process implied by the log-normal $SARV(1)$ model (12) would be able to mimic the one implied by the SR–SARV(1). However, it is problematic to justify a parametric nonlinear model through its ability to approximate features of a semiparametric linear model. Moreover Carnero, Pena and Ruiz (2004) verify that the behavior of the two autocorrelations functions can be rather different. The parametric specification tightly constrains the higher order moments in a way that may be at odds with some well documented empirical evidence, as shown in the next subsection.

At least, if one accepts the log-normal specification, it allows a rather easy computation of many nonlinear moments and thus opens the door for moment-based estimation. First proposed by Taylor (1986) and Melino and Turnbull (1990), the GMM estimation of the log-normal SARV(1) model has been thoroughly studied through Monte Carlo experiments by Andersen and Sørensen (1996). Their key observation is twofold:

First, since the assumption of Gaussianity allows an explicit computation of moments of any order, asymptotic efficiency considerations would lead to consider the largest possible set of moments to match. However, it is well known that due to fat tails in the return series, higher-order sample moments may have a very erratic finite sample behavior. This is the reason why one may be better off to match only a subset of the universe of possible moments. For simplicity Andersen and Sørensen (1996) choose to match only selected subsets of the universe of moments previously considered by Jacquier, Polson and Rossi (1994). This universe of 24 moments includes not only the aforementioned ARMA(1,1)-like structure through autocovariances between squared returns at any lag $j = 1, .., 10$ but also similar quantities for absolute returns as well as the first four unconditional moments of absolute returns.

The second key conclusion in Andersen and Sørensen (1996) is that when the efficient GMM weighting matrix is estimated more precisely and independently of the moments to match, "inclusion of additional moments almost uniformly improves estimation performance". They noted that this observa-

tion is in line with Altonji and Segal (1996) but had not at their disposal at that time the recent literature on continuously updated GMM (introduced by Hansen, Heaton and Yaron (1996)) and empirical likelihood to lessen the impact of perverse correlation between sample counterparts of the GMM efficient moment selection matrix and the moments to match. Antoine, Bonnal and Renault (2006) show that continuously updated GMM allows to annihilate such perverse correlation in the estimation of the Jacobian matrix and they propose a way to also remove this perverse correlation in the estimation of the weighting matrix. By computing higher order biases, Newey and Smith (2004) implicitly show that, in terms of higher order asymptotics, both improvements are performed automatically by empirical likelihood.

An additional difficulty comes with the fact that some matched moments are absolute values, which present a point of non-differentiability at zero, making problematic the practical computation of the Jacobian matrix. Even though this problem does not matter asymptotically because the necessary derivatives exist almost everywhere, it may be a serious concern in practice due to floating point accuracy of computer calculations. Since Melino and Turnbull (1990), there has been however a general agreement that absolute values are highly informative moments worth incorporating. Hall (2005) (p. 336–337) provides detailed practical strategies to tackle the local non differentiability problem.

2.4 Other parametric SARV models

While elementary formulas of moments of log-normal variables allowed us in the previous section to compute the autocorrelation of squared returns from the one of log-volatility assumed to be Gaussian, one may question the log-normality of volatility and rather prefer alternative transformations of the volatility process which are not as easy to deal with. Note however that autocorrelation of squared returns and autocorrelations of squared volatility are always tightly related by the general model (1). Under the simplifying assumption that the processes (σ_t) and (ε_t) are globally independent (see Section 3.2 below for a discussion of this assumption) we have for all positive lags j:

$$Cov[y_t^2, y_{t-j}^2] = Cov[\sigma_t^2, \sigma_{t-j}^2] \tag{13}$$

The autocovariances are identical, while the aucorrelations will differ by a scale factor due to a different variance. Hence, it is the decay in the autocorrelation structure that is relevant. On the maintained assumption that the squared volatility process σ_t^2 can be seen as a well defined instantaneous transformation $\sigma_t^2 = H(f_t)$ of a Gaussian and stationary process f_t, how can we compute its autocorrelation function? Granger and Newbold (1976) give an elegant solution to this problem under the very general assumption that

the function $H(\cdot)$ can be expanded in a series of Hermite polynomials. For sake of notational simplicity, let us assume without loss of generality that the Gaussian process f_t has zero mean and unit variance. Of course, in practice, this would simply warrant the introduction of two additional parameters for GMM estimation. If $(P_i, i = 0, 1, ...)$ denotes the sequence of Hermite polynomials ordered by increasing degrees, let us assume for simplicity an expansion of finite degree:

$$\sigma_t^2 = H(f_t) = \sum_{i=0}^{M} \alpha_i P_i(f_t) \tag{14}$$

The key point is that, since Hermite polynomials provide an orthogonal basis of functions for the scalar product defined by the expectation of product of functions of jointly Gaussian variables, the autocovariance function of squared volatilities is available in closed form. More precisely, knowing that if U and V are distributed as bivariate Normal with zero means, unit variances and a correlation coefficient ρ we have:

$$\begin{aligned} E[P_i(U)P_h(U)] &= 0 \qquad i \neq h \\ E[P_i(U)P_h(U)] &= i! \qquad i = h \\ E\left[P_i(U) \middle| V\right] &= \rho^i P_i(V) \end{aligned}$$

and we deduce easily:

$$Cov[\sigma_t^2, \sigma_{t-j}^2] = \sum_{i=0}^{M} \alpha_i^2 i! \rho_f^i(j)$$

where $\rho_f(j)$ denotes the correlation between f_t and f_{t-j}. Thus the autocorrelation function of the squared volatility process is given by:

$$Corr[\sigma_t^2, \sigma_{t-j}^2] = [\sum_{i=0}^{M} \alpha_i^2 i!]^{-1} \sum_{i=0}^{M} \alpha_i^2 i! \rho_f^i(j) \tag{15}$$

Of course, for the purpose of MM estimation, this can be directly related to the correlation function of the squared return process by using (13) jointly with a straightforward computation of the variance of squared returns based on kurtosis coefficients of the various components (see next section). As far as squared volatilities are concerned, (15) shows that they feature less persistence than the underlying Gaussian process as soon as the transformation is nonlinear, since $M > 1$ implies:

$$Corr[\sigma_t^2, \sigma_{t-j}^2] < \rho_f(j)$$

Overall, as developed by Meddahi(2001), equation (14), considered for a given M and parameters $(\alpha_i)_{0 \leq i \leq M}$ in some given subset of $\mathbf{R}^{M+1}$ provides a versatile parametric SARV model which generalizes the log-normal SARV model

without giving up the advantages of "linear forecasting" as in the linear SARV model. Note however that the space of parameters must be defined through a set of non trivial nonlinear constraints to ensure non-negativity of the volatility process. Meddahi (2001) actually focuses on the particular case where the "factor" f_t is an AR(1) Gaussian process ($\rho_f(j) = \rho_f^j$ with $0 < \rho_f < 1$). The key linearity property is that, thanks to the eigenfunction property of Hermite polynomials, $P_i(f_t)$ is an AR(1) process with autoregressive coefficient ρ_f^i. Since the Hermite polynomials are uncorrelated, this model can be viewed as a multifactor model, each of them being a SR–SARV(1). For instance, $P_0(f_t) + P_2(f_t) = f_t^2$ is an AR(1) process with nonnegative values and autocorrelation coefficient ρ_f^2.

An important difference relative to the SR–SARV(1) models considered in the first section is that we deal now with fully parametric models, making moment based estimation more questionable in terms of efficiency. Note however that for large M, we can theoretically capture all the functions of a latent Gaussian factor which are square integrable. Moreover, one may also consider alternative orthogonal bases of polynomials, such as the sequence of Laguerre polynomials, associated to a factor f_t driven by an AR(1) Gamma process in order to maintain the eigenfunction property (see Gouriéroux and Jasiak (2006)). In this respect, one may hope to capture all the polynomial SARV models considered by Andersen (1994) while remaining be true to the linear features of SR-SARV. This approach is thus somewhat similar in spirit to SemiNonParametric (SNP) expansions considered in Section 5 below for the purpose of efficient simulation-based estimation.

3 Implications of SV Model Specification for Higher Order Moments

3.1 Fat tails and variance of the variance

We work in this subsection under the maintained hypothesis of model (1):

$$y_{t+1} = \sigma_t \varepsilon_{t+1} \tag{16}$$

where (ε_{t+1}) is an independent white noise process, with mean zero, unit variance and kurtosis κ. The process (ε_{t+1}) as a whole is assumed to be independent of the volatility process (σ_t). This assumption will be relaxed in the next subsection dealing with the leverage effect.

3.1.1 Kurtosis, persistent volatility and volatile volatility

The kurtosis coefficient of the return process (y_{t+1}) is easily deduced from the above assumptions:

$$\frac{E(y_{t+1}^4)}{[E(y_{t+1}^2)]^2} = \kappa \left[1 + \frac{Var(\sigma_t^2)}{[E(\sigma_t^2)]^2}\right] \tag{17}$$

The decomposition (17), first put forward by Clark (1973) shows that they are two ways to accommodate well documented leptokurtic features of financial time series: either contemplating probability distributions for the standardized innovation ε with a large kurtosis coefficient κ, or specifying a volatility process (σ_t^2) with a large coefficient of variation. Of course, the decomposition is not unique. A well known example is the observational equivalence between a Student-t innovation ε and then a normal innovation combined with an additional noise term in the volatility process to remove the square-root of the chi-square induced by the denominator of the Student-t. As already noted by Rosenberg (1972): for a given observed kurtosis of returns, the kurtosis of the standardized innovation is minimized when we minimize the noise impact in the variance process. Let us now focus on how to capture a high level of kurtosis through the specification of the volatility process, for a given kurtosis κ of the standardized innovation.

Since $E(\sigma_t^2) = Var(y_t)$ is invariant to model specification, the focus of our interest is the variance of the variance $Var(\sigma_t^2)$. Consider the standard variance decomposition:

$$Var(\sigma_t^2) = Var[E\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right)] + E[Var\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right)] \tag{18}$$

It shows then, when focusing on variance predictability (or volatility persistence) to accommodate fat tails, we overlook another component. Fat tails may be accommodated through high variance $Var(\sigma_t^2)$ of the variance, not only via high volatility persistence, that is much randomness in volatility predictions $E\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right)$, but also through the second term in (18), namely a high (average) variance of the variance.

These two goals are not as distinct as they may appear. This can be illustrated by an eigenfunctions decomposition of the type (14) considered above. It implies that:

$$E\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right) = \sum_{i=0}^{M} \alpha_i \rho^i P_i(f_{t-1})$$

and thus:

$$Var[E\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right)] = \sum_{i=0}^{M} \alpha_i^2 \rho^{2i} i! \tag{19}$$

while the total variance is:

$$Var(\sigma_t^2) = \sum_{i=0}^{M} \alpha_i^2 i! \tag{20}$$

Therefore, by focusing on observed high levels of volatility persistence as measured by (19), we are tempted to put the maximum weight on the first factor ($M = 1$) since the role of additional factors $i > 1$ is dampened by exponential weighting ρ^i. Yet, it may be detrimental to neglect the benefits of additional factors ($M > 1$) in the total variance (20) of the variance. Meddahi (2001) provides evidence that the estimates reported in the literature for one-factor models typically suffer from this lack of degree of freedom and thus, cannot well accommodate fat tails.

3.1.2 Conditional moment restrictions

A practical implication of the above considerations is that, contrary to the early literature, it is worth considering regression models for capturing jointly not only fluctuations in the conditional variance (as in Section 2 above) but also fluctuations in the volatility of volatility. While general eigenfunction-based models like (14) may also be relevant in this context, we rather focus on the natural extension of the SR-SARV model introduced above by adding a quadratic specification of the conditional variance of the conditional variance:

$$Var\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right) = a + b\sigma_{t-1}^2 + c\sigma_{t-1}^4 \tag{21}$$

For simplicity, we concentrate on a $SR - SARV(1)$ model:

$$E\left(\sigma_t^2 \middle| \sigma_\tau^2, \tau < t\right) = \omega + \gamma\sigma_{t-1}^2 \tag{22}$$

The semi-parametric model defined jointly by (21) and (22) nests many popular SV models, which are characterized by constraints on the parameters (a, b, c) as well as some additional restrictions:

1. The GARCH(1,1) specification implies $a = b = 0$.
2. The Non-Gaussian Ornstein-Uhlenbeck-based SV model of Barndorff-Nielsen and Shephard (2001) implies $b = c = 0$.
3. The Heston (1993) model, a special case of the affine model of Duffie, Pan and Singleton (2000) implies $c = 0$. The value of a and b can be found in Cox, Ingersoll an Ross (1985), formula (19).

For the purpose of efficient semi-parametric estimation, it is worthwhile to apply GMM jointly to the conditional moment restrictions implied by (21) and (22). To obtain these, it helps realizing that the above specification means that the bivariate process (σ_t^2, σ_t^4) is a vector autoregressive process of order 1. In particular, the squared conditional variance process admits affine prediction formulas:

$$E\left(\sigma_t^4 \middle| \sigma_\tau^2, \tau < t\right) = a + \omega^2 + (b + 2\omega\gamma)\sigma_{t-1}^2 + (c + \gamma^2)\sigma_{t-1}^4$$

Then, from $y_{t+1} = \sigma_t \varepsilon_{t+1}$ and $E\varepsilon_{t+1}^4 = \kappa$, we deduce :

$$E\left[y_{t+1}^4 - (c + \gamma^2)y_t^4 \middle| (y_{t-j}, j \geq 1)\right] = \kappa(a + \omega^2) + \kappa(b + 2\omega\gamma)\sigma_{t-1}^2 \qquad (23)$$

It is clear from (23) that the three parameters a, b and κ cannot be identified separately. This is closely related to the observational equivalence issue described above concerning the indeterminacy of the decomposition (17). However, when plugging the linear SARV(1) dynamics (22) into (23), one obtains additional moment restrictions (see (25) below) to identify c and to improve the accuracy of the estimation of ω and γ. Basically, these additional restrictions provide relevant information about the conditional heteroskedasticity at play in the linear auto-regression equation (22). GMM estimation of the parameters ω, γ, c and ς will therefore be performed jointly on a couple of multiperiod conditional moment restrictions:

$$E\left[y_{t+1}^2 - \omega - \gamma y_t^2 \middle| (y_{t-j}, j \geq 1)\right] = 0 \qquad (24)$$

$$E\left[y_{t+1}^4 - (c + \gamma + \gamma^2)y_t^4 + \gamma(c + \gamma^2)y_{t-1}^4 \middle| (y_{t-j}, j \geq 2)\right] = \varsigma \qquad (25)$$

where ς denotes the only function of a, b and κ that we can identify. Of course, the caveat made in Subsection 2.1 is even more relevant here. The standard GMM approach works for the above set of conditional moment restrictions only if asset returns have finite moments until order eight at least. Using past returns as instruments may require further strengthening such moment conditions.

3.2 Skewness, feedback and leverage effects

Throughout Section 2, regression models for analyzing variance fluctuations were motivated by forecasting squared returns. Moreover, in line with the GARCH philosophy, conditional moment restrictions like (9) are typically used with lagged squared returns as instruments. Nelson (1991) was among the first to emphasize that there is no reason to expect the relevant conditioning information to be fully encapsulated by the past squared returns alone. In particular, the signs of past returns are informative if the relationship between past returns and future volatilities is asymmetric. Indeed, the first models of volatility clustering whereby large price movements are followed by large volatilities needed to be refined: a decline in current prices has relatively more effect on future volatilities than an increase in prices of the same size. Nelson (1991) related this stylized fact, well-documented for equity indices, to early work of Black (1976) who had attributed the

asymmetric return-volatility relationship to changes in financial leverage or debt-to-equity ratios.

Moreover focusing, as in Section 2, on volatility forecasting as distinct from return forecasting is at odds with the asset pricing literature. A non-negligible component of asset returns is a risk premium that should be related to the quantity of risk. If the latter is predictable, the former should be predictable as well. Engle, Lillien and Robbins (1987) introduced GARCH-in-Mean models with a time-varying risk premium that is a strictly increasing function of time-varying volatility. As first noted by French, Schwert and Stambaugh (1987), this approach provides a second explanation, along with the leverage effect, for the volatility asymmetry. If volatility is priced, an anticipated increase in volatility will raise the required rate of return, and necessitate an immediate asset price decline in order to allow for higher future returns. This causality from volatility to prices has been labeled the volatility feedback effect. Although it suggests a causal effect opposite to the leverage effect, which involves the reverse causality from returns to volatility, the two may be observationally equivalent if the causality lag is smaller than the time between return observations.

Bollerslev, Litvinova and Tauchen (2006) use high frequency data to try to disentangle leverage from volatility feedback effects. Relying on absolute high-frequency returns as a simple volatility proxy, their results for five-minute returns on S&P 500 futures data from the Chicago Mercantile Exchange clearly support the notion of a highly significant prolonged leverage effect at the intraday level: irrespective of the financial interpretation, there is compelling evidence that current returns and future volatilities are negatively correlated. The decay in absolute values of correlations is slow and these negative correlations remain significant for at least five days. In sharp contrast, there is little or no evidence for a delayed volatility feedback that is a negative correlation between current volatility and future returns. However, since the correlation between current returns and current volatility is even more strikingly negative than with respect to future volatilities, the authors conclude that this can be interpreted as evidence in favor of an instantaneous volatility feedback effect.

However, an alternative interpretation of the Bollerslev, Litvinova and Tauchen (2006) results may be found in the conditional skewness of asset returns. To illustrate the potential role for conditional skewness, let us consider the extreme case where there is no risk premium ($m_t = 0$) in returns. Then the conditional covariance at time t between the next return and its squared value can be written as:

$$Cov_t[y_{t+1}^2, y_{t+1}] = \sigma_t^3 E_t[\varepsilon_{t+1}^3]$$

In other words, the observed strongly negative correlation between return and its contemporaneous volatility may be due to negative conditional skewness. Additional work is needed to disentangle the two effects. Finally, it is worth

noting that such negative skewness is implied by a leverage effect at higher frequencies. Let us imagine for instance, still in the zero risk premium case, a negative correlation between current return and future squared returns as typically implied by leverage effect:

$$Cov_t[y_{t+1}, \sigma^2_{t+1}] < 0 \Longrightarrow Cov_t[y_{t+1}, y^2_{t+2}] < 0$$

Then, with a twice lower frequency, the conditional skewness of returns will be computed as:

$$E_t[(y_{t+1} + y_{t+2})^3] = Cov_t[(y_{t+1} + y_{t+2}), (y_{t+1} + y_{t+2})^2]$$

and may be negative simply due to leverage while there is no skewness at the higher frequency:

$$E_t[(y_{t+1})^3] = Cov_t[y_{t+1}, y^2_{t+1}] = 0$$
$$E_t[(y_{t+2})^3] = Cov_t[y_{t+2}, y^2_{t+2}] = 0$$

As far as moment-based inference is concerned, the conditional moment restrictions introduced in former sections may be completed respectively by:

$$E_t[(y_{t+1})(y^2_{t+2})] = 0$$

and/or by:

$$E_t[(y^3_{t+1})] = 0$$

to respectively test (or account) for zero leverage and/or zero skewness.

4 Continuous Time Models

We emphasized in Section 2 that both the weak-GARCH and the SR-SARV model are robust to time aggregation. Even more interestingly, as respectively shown by Drost and Werker (1996) and Meddahi and Renault (2004), they can be seen as resulting from discrete time sampling in a continuous time diffusion model with stochastic volatility, insofar as the factors spanning the stochastic volatility process have a linear drift. We will consider more generally in this section a typical continuous time SV model for log-prices $p(t)$ with stochastic volatility and finite jump activity:

$$dp(t) = \mu(t)\, dt + \sigma(t)\, dW(t) + \kappa(t)\, dq(t) \tag{26}$$

where $\mu(t)$ is a continuous and locally bounded variation process and $\sigma(t)$ is a strictly positive and càdlàg stochastic volatility process and $W(t)$ is a Wiener process, $dq(t)$ is a counting process with $dq(t) = 1$ corresponding

to a jump at t and $dq(t) = 0$ if no jump. The (possibly time-varying) jump intensity is $\lambda(t)$ and $\kappa(t)$ is the jump size.

There is a vast literature on the estimation of continuous time SV models. We will therefore be selective in our coverage. In this section we focus on moment-based estimation involving measures of realized volatility while simulation-based MM estimators will be described in the next section.

4.1 Measuring volatility

The volatility measure appearing in equation (28) is not observable but can be estimated from data. Estimation, which comes with sampling error, is based on the increment of the quadratic variation over some horizon H, that is

$$QV_{t,t+H} = \int_t^{t+H} \sigma^2(s)\, ds + \sum_{\{s\in[t,t+H]:dq(s)=1\}} \kappa^2(s). \tag{27}$$

Note that in the absence of jumps, the quadratic variation equals the integrated variance over the period H, namely $\sigma^{[2]}_{t,t+H}$ defined as:

$$\sigma^{[2]}_{t,t+H} = \int_t^{t+H} \sigma^2(s)\, ds. \tag{28}$$

There is now a well established asymptotic theory which pertains to statistics based on samples over finite intervals involving data observed at ever increasing frequency (see Jacod (1994, 1997), Barndorff-Nielsen and Shephard (2002) as well as the recent survey by Barndorff-Nielsen and Shephard (2007)). To proceed with estimation we define the (discrete) daily log return as $r_{t,t-1} = \ln P_t - \ln P_{t-1} = p_t - p_{t-1}$ where the t refers to daily sampling (henceforth we will refer to the time index t as daily sampling). The intra-daily return is then denoted $r_{t,t-1/M} = p_t - p_{t-1/M}$ where $1/M$ is the (intra-daily) sampling frequency. It is possible to consistently estimate $QV_{t,t+H}$ in (27) by summing squared intra-daily returns, yielding the so-called realized variance, namely:

$$RV^M_{t,t+H} = \sum_{j=1}^{MH} \left(r_{(t+H)-j/M,(t+H)-(j-1)/M}\right)^2. \tag{29}$$

To facilitate the presentation we simplify notation by focusing exclusively on one-day ($H = 1$) quadratic variation and related measures which will be introduced shortly. Moreover, we will henceforth drop the superscript M. Hence, we write $RV_{t,t+1}$ instead of $RV^M_{t,t+1}$ and let all limits and convergence in distribution arguments apply to $M \to \infty$ although M does not explicitly appear in the simplified notation. Jacod (1994, 1997) and Barndorff-Nielsen

and Shephard (2002) show that the error of realized variance is asymptotically

$$\frac{RV_{t,t+1} - QV_{t,t+1}}{\sqrt{2\sigma^{[4]}_{t,t+1}/M}} \sim N(0,1) \tag{30}$$

where $\sigma^{[4]}_{t,t+1} = \int_t^{t+1} \sigma(s)^4 ds$. In the absence of jumps a feasible asymptotic distribution is obtained by replacing $\sigma^{[4]}_t$ with a sample equivalent, namely,

$$RQ_{t,t+1} = (M/3) \sum_{j=1}^{M} \left(r_{t-j/M,t-(j-1)/M}\right)^4 \tag{31}$$

which is called the realized quarticity. In the presence of jumps various alternative measures of volatility like bipower variation have been proposed, see the recent survey by Barndorff-Nielsen and Shephard (2007) for further discussion.

Early work by Andreou and Ghysels (2002) as well as a recent contribution of Ghysels, Mykland and Renault (2007) have noted that while in the limit in-sample observations suffice to estimate current realized variation, there are efficiency gains for any finite sample configuration, that is, there are gains to be made in practical applications of extracting realized volatility to use realized volatility from the previous days. This means that measurement of volatility and modeling of volatility are mutually intertwined since refined measures of volatility involve prediction models with possibly a price to pay in terms of additional assumptions. Moreover, when it comes to prediction, it is quite important to prefilter non-persistent shocks (jumps) thanks to alternative volatility measurements like bipower variation (see e.g. Andersen, Bollerslev and Diebold (2007)).

4.2 Moment-based estimation with realized volatility

There is a substantial literature on the estimation of SV models using (daily) returns data. Can we possibly estimate stochastic volatility models such as in (26) using realized volatility. The answer is affirmative, at least if we restrict attention to specific models allowing for closed-form solutions to the moments of volatility.

Let us first consider for the sake of notational simplicity a linear SARV model with only one factor. In continuous time, such a model is generally defined as:

$$\begin{aligned} dp(t) &= \sigma(t) dW_1(t) \\ d\sigma^2(t) &= k(\theta - \sigma^2(t))dt + \xi(t) dW_2(t) \end{aligned} \tag{32}$$

where $(W_1(t), W_2(t))'$ is a bivariate Brownian motion ($W_1(t)$ and $W_2(t)$ are standard Brownian motions which are possibly correlated). Note that we don't need to be specific about the diffusion term $\xi(t)$ for the volatility process. The key moments conditions below (35) and (37) only depend on the linear drift $k(\theta - \sigma^2(t))$. Only when it comes to the variance of the variance do we need to be more specific about a functional dependence between the process $\xi(t)$ and the spot volatility process $\sigma(t)$.

Meddahi and Renault (2004) show that (32) implies that the daily return process $y_{t+1} = r_{t,t+1} = p_{t+1} - p_t$ is a SR–SARV(1). More precisely, $y_{t+1} = \sigma_t \varepsilon_{t+1}$with:

$$\sigma_t^2 = (1/k)(1 - e^{-k})\sigma^2(t) + [1 - (1/k)(1 - e^{-k})]\theta \tag{33}$$

σ_t^2 inherits the linear autoregressive structure implied by the drift in (32). It implies a linear autoregression for volatility as in equation (22) with $\gamma = e^{-k}$ and $\omega = \theta(1 - e^{-k})$. From daily return data, this allows for estimation of the parameters k and θ via the conditional moment restrictions (24). Higher orders p of volatility autoregressive dynamics (and associated MM estimation based on (9) with $q = p$) would similarly appear when considering p factors with linear drifts within a diffusion setting like (32). It takes the form of a simple generalization of (33) with σ_t^2 expressed as an affine function of the p continuous time factors (see Meddahi and Renault (2004), Prop 2.1.).

However, with high-frequency intraday data, Bollerslev and Zhou (2002) have proposed using time series of daily realized variances directly for GMM estimation. The key is that σ_t^2 represents the conditional expectation at time t of the integrated variance $\sigma_{t,t+1}^{[2]}$. Thus, with the common notation E_t for the conditional expectation operator at time t, equation (22) simply tells us that:

$$E_{t-1}[E_t(\sigma_{t,t+1}^{[2]})] = \omega + \gamma E_{t-1}(\sigma_{t-1,t}^{[2]})$$

or by the Law of Iterated Expectations:

$$E_{t-1}[\sigma_{t,t+1}^{[2]} - \omega - \gamma\sigma_{t-1,t}^{[2]}] = 0 \tag{34}$$

which is exactly the conditional moment restriction (6) of Bollerslev and Zhou (2002). To get a moment restriction based on observed realized variances, Bollerslev and Zhou (2002) noted that, due to the zero drift assumption in log-price $p(t)$, realized variance $RV_{t,t+1}$ is not only a consistent but also a "E_t-unbiased" estimator of integrated variance $\sigma_{t,t+1}^{[2]}$. Therefore, equation (34) is also valid for realized variances:

$$E_{t-1}[RV_{t,t+1} - \omega - \gamma RV_{t-1,t}] = 0 \tag{35}$$

Bollerslev and Zhou (2002) consider also more generally the case where the spot squared volatility process $\sigma^2(t)$ is the sum of two processes like above:

$$\begin{aligned} \sigma^2(t) &= V_1(t) + V_2(t) \\ dV_1(t) &= k_1(\theta_1 - V_1(t))dt + \xi_1(t)dW_{21}(t) \\ dV_2(t) &= k_2(\theta_2 - V_2(t))dt + \xi_2(t)dW_{22}(t) \end{aligned} \tag{36}$$

As shown by Meddahi and Renault (2004), this two-factors volatility model implies a SR–SARV(2) model for daily returns, again thanks to the linearity of the drift of the volatility factors. Again, from the ARMA(2,2) structure of squared returns in the SR–SARV(2) case (conditional moment restriction based on (9) with $q = p = 2$), Bollerslev and Zhou (2002) (see their appendix B) are able to deduce similar moment conditions extending (35) to the two-factors case. They are even able to accomodate a Poisson jump component with constant intensity and log-normal jump size. Consistent with other recent findings in the literature, their empirical results suggest the presence of multiple volatility factors and/or jumps. Note that, irrespective of the number of factors, doing MM estimation from (35) as Bollerslev and Zhou (2002) do with five-minute returns data, should be more accurate than the MM estimators based on (24), because between two "E_t -unbiased" estimator of integrated variance $\sigma^{[2]}_{t,t+1}$, namely y^2_{t+1} and $RV_{t,t+1}$, the latter is least volatile (less noisy). Note also that, for instance in the one-factor case, an alternative way to use the information brought by five-minute returns would have been to directly apply (24) to five-minute returns:

$$E_{t-k\delta}[r^2_{t-(k-1)\delta,t-(k-2)\delta} - \omega(\delta) - \gamma(\delta) r^2_{t-k\delta,t-(k-1)\delta}] = 0 \tag{37}$$

with $\omega(\delta) = e^{-k\delta}$ and $\gamma(\delta) = \theta(1 - e^{-k\delta})$. Under the maintained assumption that the continuous time model (32) is well-specified, the MM estimator of the structural parameters k and θ obtained from (37) makes more efficient use of the available information (five-minute returns) than the one derived from (35). Why then rather using (35)? Three kinds of explanations may be put forward.

1. High-frequency data are delicate to use directly, due to many kinds of potential problems, including measurement errors, microstructure noise, missing data and the cost of manipulating massive time series.
2. High-frequency data involve irregular and possibly random times of sampling.
3. Last but not least, five-minute returns are hardly informative about the autocorrelation patterns of squared returns, because the informative content of moment conditions like (37) is blurred by very strong intraday seasonality in volatility. Andersen and Bollerslev (1997) bring compelling evidence that intraday seasonality destroys the autocorrelation patterns in absolute/squared returns when high-frequency data are used, thus producing non-sensical parameter estimates for many studies estimating typical ARCH or SV models.

However, explanations (i) and/or (iii) leave open the issue of choosing the best aggregator, instead of focusing exclusively on daily realized variances. Explanation (ii) motivates explicitly taking into account the presence of irregular and random arrival times of the return data in the moment conditions. While the aggregator choice may in particular incorporate features of the intraday seasonality, a direct empirical study of volatility dynamics from tick by tick data providing quote changes at random times must incorporate a time-of-day dummy (see e.g. Renault and Werker (2008)). While a more thorough discussion of intraday seasonality is beyond the scope of this chapter, the two issues of daily aggregation of intraday data and random sampling times for volatility measurement will be respectively addressed in Subsections 4.3 and 4.4 below.

Bollerslev and Zhou (2002) go one step further than MM based on (35) by using not only the linear SARV structure (22) provided by the linear drift in (32) but also a Heston (1993) square root specification of the volatility process, that is $\xi^2(t)$ is an affine function of $\sigma^2(t)$. As already mentioned in Section 2, these joint assumptions provide a linear vectorial autoregressive structure for (σ_t^2, σ_t^4) conformable to (21) and (22). As stated in Section 2 $E_t(\sigma_{t+1}^4)$ is an affine function of σ_t^2 and σ_t^4, and Bollerslev and Zhou (2002) obtain a similar result for $E_t[(\sigma_{t,t+1}^{[2]})^2]$. More precisely, it is an affine function of $\sigma^2(t)$ and $\sigma^4(t)$ by their formula (A7), and in turn of σ_t^2 and σ_t^4 by applying (33) above again. The key issue is then to deduce a conditional moment restriction about observed returns. While this had been made possible in Section 2 by assuming a constant conditional kurtosis, Bollerslev and Zhou (2002) circumvent the difficulty by proceeding as if squared realized variance $(RV_{t,t+1})^2$ was an E_t-unbiased estimator of squared integrated variance $(\sigma_{t,t+1}^{[2]})^2$. Since $E_t[RV_{t,t+1}] = E_t[\sigma_{t,t+1}^{[2]}]$, this approximation requires that the conditional variances are near identical:

$$V_t[RV_{t,t+1}] \approx V_t[\sigma_{t,t+1}^{[2]}] \tag{38}$$

Sidestepping technicalities, this approximation is applicable for a large number of intraday observations, M, due to the convergence result (30). First, it is worth knowing that this convergence is stable (see Jacod and Shiryaev (1987), Chap. 8) which, intuitively, allows us to use the convergence result for moments involving jointly the converging sequence and functions of the volatility process. Then, we can heuristically deduce from (30) that:

$$V_t[RV_{t,t+1}] = V_t[\sigma_{t,t+1}^{[2]}] + (2/M)E_t[\sigma_{t,t+1}^{[4]}] + o(1/M) \tag{39}$$

or almost equivalently:

$$V_t[RV_{t,t+1}] = V_t[\sigma_{t,t+1}^{[2]}] + (2/M)E_t[RQ_{t,t+1}] + o(1/M) \tag{40}$$

A formal derivation of (39) and (40) from (30) requires showing that the (conditional) covariance between $\sigma^{[2]}_{t,t+1}$ and the squared asymptotic standard normal error is of order less than $(1/M)$ (see also Andersen, Bollerslev and Meddahi (2005) for an unconditional version of (40)). Corradi and Distaso (2006) recently put forward a more formal analysis in a double asymptotic setting. Still, they only consider unconditional moments.

In order to compare with the Section 2 approach of MM directly on asset returns ($M = 1$) rather than on realized variances (with a large M), note that with the assumptions of Section 2, we would have:

$$V_t[RV_{t,t+1}] = V_t[y^2_{t+1}] = (\kappa - 1)\sigma^4_t$$

while:

$$(2/M)E_t[RQ_{t,t+1}] = (2/3)E_t[y^4_{t+1}] = (2/3)\kappa\sigma^4_t$$

In other words, not only the term $(2/M)E_t[RQ_{t,t+1}]$ would not be a negligible error in the difference between $V_t[RV_{t,t+1}]$ and $V_t[\sigma^{[2]}_{t,t+1}]$ but it would coincide with the former in case of conditional normality of returns ($\kappa = 3$). As already explained, high-frequency data allow us to obtain an unbiased estimator, $RV_{t,t+1}$ of $\sigma^{[2]}_{t,t+1}$ which is much less volatile than the daily squared return and in the limit has no more volatility than $\sigma^{[2]}_{t,t+1}$. This is achieved by cumulating, say, M equally-spaced intraday return observations. Intuitively, this enables us to divide the spread $2E_t[RQ_{t,t+1}]$ by a factor of M. The Bollerslev and Zhou (2002) approximation is then clearly valid as they apply it with five-minutes returns. In a similar vein Garcia, Lewis, Pastorello and Renault (2006) add moment conditions based on the third conditional moment of integrated volatility as it helps to better identify the asymmetry and leverage effects. They also use option price data via the Heston (1993) option pricing model. Bollerslev, Gibson and Zhou (2004) adopted a very similar approach, but considered a so-called model-free approach to recover implied volatilities. There is clearly a trade-off between model-free and model-based approaches to recover implied volatilities. While a model-free approach is robust to misspecification, it requires theoretically continuous strikes for option prices or practically a very liquid market like the S & P 500 option market. In contrast, model-based approaches may be sensitive to misspecification but they require only a few option prices.

4.3 Reduced form models of volatility

The most successful volatility forecasting models before the advent of high frequency data were ARCH-type models. These are roughly speaking time series models applied to squared returns. It is therefore also not surprising that today the most successful models are time series models applied to realized

volatility models. These are pure forecasting models, and therefore do not relate to any explicit diffusion model. The bulk of these models use daily aggregates ranging from realized volatility appearing in equation (29) or similar measures that separate jumps from the continuous component. These models are reviewed in Andersen et al. (2006) and Barndorff-Nielsen and Shephard (2007).

Volatility forecasting typically involves predictions over multiple horizons, e.g., via Value-at-Risk computations for risk management purposes, whereas the data is sampled potentially at intra-daily frequency. An approach particularly adept at handling such situations is the the mixed data sampling, or MIDAS approach of Ghysels, Santa-Clara and Valkanov (2005). There have been many successful applications in the context of volatility forecasting, including Forsberg and Ghysels (2006), Ghysels, Santa-Clara and Valkanov (2006) and Ghysels and Sinko (2006). A generic MIDAS regression model is as follows:

$$RV_{t,t+H}^{M} = a + b\sum_{j=1}^{\tau} \psi_j(\theta) X_{t-j} + \varepsilon_t \tag{41}$$

for various regressors X and the parameters are a function of a small set of hyperparameters θ. Various polynomial specifications are discussed in Ghysels, Sinko and Valkanov (2006). The regressors can be sampled at any frequency, not necessarily daily, and the horizon is also arbitrary. The most powerful predictive regressors usually involve absolute returns, such as realized absolute values, as documented by Forsberg and Ghysels (2006). Chen and Ghysels (2008) provide an example of a semi-parametric MIDAS regression where X is replaced by an unknown function of high frequency returns, i.e. $m(r_{t-j/M})$ and a polynomial that captures both daily and intra-daily decay patterns.

4.4 High frequency data with random times separating successive observations

In this section we assume that the asset price P_t is observed at irregularly spaced dates $t_0, t_1, ..., t_n$, with $0 = t_0 < t_1 < ... < t_n$. We denote by $\delta_i, i = 1, ..., n$ the ith duration ($\delta_i = t_i - t_{i-1}$) and, for sake of notational simplicity, we assume that returns have a zero conditional mean. We also assume that we control for intraday volatility patterns, either via an intraday seasonal correction factor, or by focusing on the same given time interval every day.

In his simplest volatility model, Engle (2000) assumes that the variable σ_{i-1}^2 defined as:

$$\sigma_{i-1}^2 = \frac{h_{i-1}}{\delta_i}$$

where $h_{i-1} = Var[y_{t_i}|y_{t_j}, \delta_j, j \leq i-1; \delta_i]$ follows a GARCH(1,1)-type equation. More precisely, under the assumption:

$$E[y_{t_i}|y_{t_j}, \delta_j, j \leq i-1; \delta_i] = 0 \tag{42}$$

Engle (2000) specifies:

$$\sigma_{i-1}^2 = \omega + \alpha(y_{t_{i-1}}/\sqrt{\delta_{i-1}})^2 + \beta\sigma_{i-2}^2 \tag{43}$$

In other words, in order to take into account the unequally spaced feature of the returns, Engle (2000) assumes that the variance per time unit σ_i^2 follows a GARCH(1,1) equation. In contrast, Ghysels and Jasiak (1998) specify a GARCH equation for the total variance process h_i^* defined by;

$$h_{i-1}^* = Var[y_{t_i}|y_{t_j}, \delta_j, j \leq i-1] \tag{44}$$

However, in order to take into account the unequally spaced nature of the returns, Ghysels and Jasiak (1998) assume a time-varying parameter GARCH equation with:

$$h_{i-1}^* = \omega_{i-1} + \alpha_{i-1}(y_{t_{i-1}})^2 + \beta_{i-1}h_{i-2}^* \tag{45}$$

where the parameters ω_{i-1}, α_{i-1}, β_{i-1} are functions of the expected duration $\Psi_{i-1} = E[\delta_i|y_{t_j}, \delta_j, j \leq i-1]$. The functional forms adopted by Ghysels and Jasiak (1998) are inspired by the weak GARCH representation of a GARCH diffusion model put forward by Drost and Werker (1996) in the case of equally spaced observation ($\delta_i = \delta \ \forall \ i$). More precisely, Ghysels and Jasiak (1998) postulate an extended validity of temporal aggregation formulas for weak GARCH by assuming:

$$\alpha_i + \beta_i = exp(-k\Psi_i) \ .$$

It is worth stressing several differences between Engle (2000) and Ghysels and Jasiak (1998). In contrast with Ghysels and Jasiak (1998), Engle (2000) includes the current duration in the conditioning information. By the law of iterated expectation, h_{i-1}^* is only the best forecast of h_{i-1} given past returns and durations.

To better illuminate the differences between the two approaches, it is worth revisiting the exact discretization of the continuous time model (32). Since for any given duration δ :

$$E[\sigma^2(t+\delta)|\sigma^2(\tau, \tau \leq t)] = \theta + exp(-k\delta)(\sigma^2(t) - \theta)$$

A simple computation shows that when the processes W, σ and the durations are mutually independent:

$$Var[y_{t_i}|y_{t_j}, \delta_j, j \leq i-1; \delta_i] = \theta\delta_i + c(k\delta_i)(\sigma_{i-1}^2 - \theta)\delta_i$$

where $c(\delta) = (1/\delta)[1 - exp(-\delta)]$. For small durations, the function $c(\cdot)$ is almost one, which shows that, in the line of Engle (2000) it is rather natural to focus on the conditional variance by unit of time h_{i-1}. Meddahi, Renault

and Werker (2006) show that in this setting:

$$h_{i-1} = \omega_i + \gamma_i h_{i-2} + \eta_{i-1}$$

where

$$\gamma_i = exp(-k\delta_{i-1})\frac{c(k\delta_i)}{c(k\delta_{i-1})}, \quad \omega_i = \theta(1-\gamma_i), \quad E[\eta_{i-1}|y_{t_j}, \delta_j, j \leq i-1; \delta_i] = 0\,.$$

Therefore, for the volatility per unit of time defined as in Engle (2000), we end up with something like an AR(1) structure of volatility, conformable to the spirit of SR-SARV(1), but with a time varying autoregressive coefficient along the lines of Ghysels and Jasiak (1998). In terms of conditional moment restrictions valid for inference on the structural parameters of interest, Meddahi, Renault and Werker (2006) put forward the following generalization of (24):

$$E[(y_{t_i}/\sqrt{\delta_i})^2 - \omega_i - \gamma_i(y_{t_{i-1}}/\sqrt{\delta_{i-1}})^2|y_{t_j}, \delta_j, j \leq i-2] = 0$$

Recall, however, that such restrictions are valid only when the durations are assumed to be independent from the volatility process. Renault and Werker (2008) show that, more often than not, there is a significant negative correlation between volatility and current duration. As a result, the conditional volatility of future return given past information and current duration is significantly smaller than the spot volatility multiplied by expected duration. Renault and Werker (2008) propose a setting to extend the above conditional moment restrictions to such a case of endogenous durations. One complication is that the expectation of functions of durations like $c(\cdot)$ above involves the Laplace transform of the conditional distribution of durations. In other words, a semi-parametric moment approach is more difficult to maintain. Renault and Werker (2008) circumvent this difficulty by assuming that with high frequency data, the spot volatility process can be viewed as a martingale.

5 Simulation–Based Estimation

For a parametric stochastic volatility model, the simulation tool and Monte Carlo integration provide a versatile minimum distance estimation principle well suited to accommodate latent volatility factors. The general approach is labeled Simulation-based Indirect Inference (SII). It can take advantage of any instrumental piece of information that identifies the structural parameters. Examples include the Simulated Method of Moments (SMM) and its simulated-score matching version, leading to the so-called Efficient Method of Moments (EMM). However, since the simulator is governed by the struc-

tural model, the classical trade-off between efficiency and robustness should be revisited.

5.1 Simulation-based bias correction

Let θ denote a vector of p unknown parameters. We want to build an accurate estimator $\widetilde{\theta}_T$ of θ from an observed sample path of length T. Let us assume that we have at our disposal some initial estimator, denoted by $\widetilde{\beta}_T$. Note that we use on purpose a letter β different from θ to stress that the estimator $\widetilde{\beta}_T$ may give a very inaccurate assessment of the true unknown θ^0 it is supposed to estimate. In particular this estimator is possibly severely biased: its expectation $b_T(\theta^0)$ does not coincide with θ^0. The notation $b_T(\theta^0)$ indicates that the so-called binding function (Gouriéroux and Monfort (1992)) depends on at least two things: not only on the true unknown value of the parameters of interest but also on the sample size. We will consider here a bias correction procedure that can be seen as a form of parametric bootstrap performed in a non-linear state space model of the general form:

$$\begin{aligned} y_t &= r_1[y(0,t-1), y^*(0,t), \varepsilon_{1t}, \theta], t = 1, ..., T \\ y_t^* &= r_2[y(0,t-1), y^*(0,t-1), \varepsilon_{2t}, \theta], t = 1, ..., T \end{aligned} \tag{46}$$

where $\varepsilon_t = (\varepsilon_{1t}', \varepsilon_{2t}')'$ is a white-noise process whose marginal distribution P_ε is known, (y_t^*) is a process of latent variables, typically latent volatility factors, and $r_1$and r_2 are two known functions. The parametric model (46) recursively defines the observed endogenous variables through a path $y^*(0,T) = (y_t^*)_{0 \le t \le T}$ of latent ones, making path simulations possible. More precisely, from independent random draws ε_t^h, $t = 1, ...T$ and $h = 1, ..., H$ in P_ε, we can now compute recursively:

$$y_t^{*h}(\theta) = r_2[y^h(0,t-1)(\theta), y^{*h}(0,t-1)(\theta), \varepsilon_{2t}^h, \theta], t = 1, ..., T; h = 1, ..., H$$

$$y_t^h(\theta) = r_1[y^h(0,t-1)(\theta), y^{*h}(0,t)(\theta), \varepsilon_{1t}^h, \theta], t = 1, ..., T; h = 1, ..., H$$

Note that, due to the presence of a dynamic process of latent state variables, the draw of $y_t^h(\theta)$ at each given t is conditional on past simulated $y^h(0,t-1)(\theta)$ and not on past observed $y(0,t-1)$. Hence the terminology path simulations (see Gouriéroux and Monfort (1996)). In the spirit of parametric bootstrap, that is resampling from a preliminary estimator, $\widetilde{\beta}_T$ gives rise to H bootstrap samples $y^h(0,T)(\widetilde{\beta}_T), h = 1, ..., H$. On each bootstrap sample, the same estimation procedure can be applied to get H estimators denoted as $\beta_T^h(\widetilde{\beta}_T), h = 1, ..., H$. These estimations characterize the bootstrap distribution of $\widetilde{\beta}_T$ and allow for instance to approximate the unknown expectation $b_T(\theta^0)$ by $b_T(\widetilde{\beta}_T)$. Of course, $b_T(\widetilde{\beta}_T)$ is not known in general but may be approximated at any desired level of accuracy, from the Monte

Carlo average $\frac{1}{H}\sum_{h=1}^{H}\beta_T^h(\widetilde{\beta}_T)$, for H sufficiently large. The bias-corrected bootstrap estimator is then defined as:

$$\widetilde{\theta}_T = \widetilde{\beta}_T - [b_T(\widetilde{\beta}_T) - \widetilde{\beta}_T]\,. \tag{47}$$

However, this parametric bootstrap procedure hinges on a sensible initial estimator, $\widetilde{\beta}_T$, so that we are confident that the estimated bias $[b_T(\widetilde{\beta}_T) - \widetilde{\beta}_T]$ provides a good assessment of the true bias $[b_T(\theta^0) - \theta^0]$. For this reason Gouriéroux, Renault and Touzi (2000) propose an alternative iterative procedure which, at step j, improves upon an estimator $\widetilde{\theta}_T^j$ by computing $\widetilde{\theta}_T^{j+1}$ as:

$$\widetilde{\theta}_T^{j+1} = \widetilde{\theta}_T^j + \lambda[\widetilde{\beta}_T - b_T(\widetilde{\theta}_T^j)] \tag{48}$$

for some given updating parameter λ between 0 and 1. In other words, at each step, a new set of simulated paths $y^h(0,T)(\widetilde{\theta}_T^j), h = 1, ..., H$, is built and it provides a Monte Carlo assessment $b_T(\widetilde{\theta}_T^j)$ of the expectation of interest. It is worth reminding that this does not involve new random draws of the noise ε. Note that (47) corresponds to the first iteration of (48) in the particular case $\lambda = 1$ with a starting value $\widetilde{\theta}_T^1 = \widetilde{\beta}_T$. While this preliminary estimator is indeed a natural starting value, the rationale for considering λ smaller than 1 is to increase the probability of convergence of the algorithm, possibly at the cost of slower convergence (if faster update would also work). Incidently, if this algorithm converges, the limit defines an estimator $\widetilde{\theta}_T$ which solves,

$$b_T(\widetilde{\theta}_T) = \widetilde{\beta}_T\,. \tag{49}$$

Gouriéroux, Renault and Touzi (2000) study more generally the properties of the estimator (49) which represents a special case of SII estimators developed in the next subsection. The intuition is quite clear. Let us call $\widetilde{\beta}_T$ the naive estimator. Our preferred estimator $\widetilde{\theta}_T$ is the value of unknown parameters θ, which, if it had been the true one, would have generated a naive estimator which, on average, would have coincided with our actual naive estimator. In particular, if the bias function $[b_T(\theta) - \theta]$ is linear with respect to θ, we deduce $b_T[E(\widetilde{\theta}_T)] = E(\widetilde{\beta}_T) = b_T(\theta^0)$ and thus our estimator is unbiased. Otherwise, unbiasedness is only approximately true to the extent a linear approximation of the bias is reasonable. Noting that in the context of stationary first order autoregressive processes, the negative bias of the OLS estimator of the correlation coefficient becomes more severely non-linear in the near unit root case, Andrews (1993) proposed a median-unbiased estimator based on the principle (49) with median replacing expectation. The advantage of the median is that it is immune to non-linear monotonic transformations, while the drawback is that it is hard to generalize to a multi-parameter setting. As for linear autoregressive processes, estimation of SARV models with highly persistent volatility may result in a significant downward bias in the estimates of volatility persistence. Pastorello, Renault and Touzi (2000) document how

the bias correction procedure is useful for volatility models with (log)linear drift as specified in prior sections.

5.2 Simulation-based indirect inference

The key intuition of indirect inference is that defining an indirect estimator of the parameters of interest via an initial estimator $\widetilde{\beta}_T$ and a binding function $b_T(\cdot)$ by solving the equation:

$$b_T(\widetilde{\theta}_T) = \widetilde{\beta}_T$$

is worthwhile beyond the bias-correction setting studied above. The vector β of so-called instrumental parameters must identify the structural parameters θ but does not need to bear the same interpretation. In the early linear simultaneous equations literature, an example of indirect inference was put forth under the label "Indirect Least Squares" : the instrumental parameters β, the coefficients of the reduced form, are estimated by OLS, while solving equation (49) provides a consistent estimator of the structural parameters θ. However, this historical example is too simple to display all the features of Indirect Inference as more generally devised by Smith (1993) and Gouriéroux, Monfort and Renault (1993) for two reasons:

First, the binding function is not in general available in closed form and can be characterized only via Monte Carlo integration. Moreover, by contrast with the simple linear example, the binding function, in general, depends on the sample size T.

Second, most interesting examples allow for overidentification of the structural parameters, for instance through a multitude of instrumental variables in the simultaneous equation case. This is the reason why we refer henceforth to the auxiliary parameters β as instrumental parameters and assume that the dimension of β is larger than (or equal to) the one of θ.

The key idea is that, as already explained in the former subsection, our preliminary estimation procedure for instrumental parameters not only gives us an estimation $\widetilde{\beta}_T$ computed from the observed sample path but also can be applied to each simulated path $y^h(0,T)(\theta), h = 1, ..., H$. Thus, we end up, possibly for each value of θ, with a set of H "estimations"$\beta_T^h(\theta), h = 1, ..., H$. Averaging them, we get a Monte Carlo binding function:

$$\beta_{T,H}(\theta) = \frac{1}{H}\sum_{h=1}^{H} \beta_T^h(\theta)\,.$$

The exact generalization of what we did in the previous subsection amounts to define the binding function $b_T(\theta)$ as the probability-limit (w.r.t. the random draw of the process ε) of the sequence $\beta_{T,H}(\theta)$ when H goes to infinity.

However, for most non-linear models, the instrumental estimators $\beta_T^h(\theta)$ are not really reliable for finite T but only for a sample size T going to infinity. It is then worth realizing that when T goes to infinity, for any given $h = 1, ..., H$, $\beta_T^h(\theta)$ should tend towards the so-called asymptotic binding function $b(\theta)$ which is also the limit of the finite sample binding function $b_T(\theta)$.

Therefore, as far as consistency of estimators when T goes to infinity is concerned, a large number H of simulations is not necessary and we will define more generally an indirect estimator $\widetilde{\theta}_T$ as solution of a minimum distance problem:

$$\min_{\theta}[\widetilde{\beta}_T - \beta_{T,H}(\theta)]' \Omega_T [\widetilde{\beta}_T - \beta_{T,H}(\theta)] \tag{50}$$

where Ω_T is a positive definite matrix converging towards a deterministic positive definite matrix Ω. In case of a completed Monte Carlo integration (H large) we end up with an approximation of the exact binding function-based estimation method:

$$\min_{\theta}[\widetilde{\beta}_T - b_T(\theta)]' \Omega_T [\widetilde{\beta}_T - b_T(\theta)] \tag{51}$$

which generalizes the bias-correction procedure of the previous subsection. As above, we may expect good finite sample properties of such an indirect estimator since, intuitively, the finite sample bias is similar in the two quantities which are matched against each other and thus should cancel out through the differencing procedure.

In terms of asymptotic theory, the main results under standard regularity conditions (see Gouriéroux, Monfort and Renault (1993)) are:

1. The indirect inference estimator $\widetilde{\theta}_T$ converges towards the true unknown value θ^0 insofar as the asymptotic binding function identifies it:
$$b(\theta) = b(\theta^0) \Longrightarrow \theta = \theta^0 .$$
2. The indirect inference estimator $\widetilde{\theta}_T$ is $\sqrt{T}$ - asymptotically normal insofar as the asymptotic binding function first-order identifies the true value via a full-column rank for:
$$\frac{\partial b}{\partial \theta'}(\theta^0) .$$
3. We get an indirect inference estimator with a minimum asymptotic variance if and only if the limit-weighting matrix Ω is proportional to the inverse of the asymptotic variance Σ_∞ of $\sqrt{T}[\widetilde{\beta}_T - b_T(\theta^0)]$.
4. The asymptotic variance of the efficient indirect inference estimator is the inverse of $[\frac{\partial b'}{\partial \theta}(\theta^0)(\Sigma_H)^{-1}\frac{\partial b}{\partial \theta'}(\theta^0)]$ with $\Sigma_H = (1 + \frac{1}{H})\Sigma_\infty$.

An implication of these results is that, as far as asymptotic variance of the indirect inference estimator is concerned, the only role of a finite number H of simulations is to multiply the optimal variance (obtained with $H = \infty$) by a factor $(1 + \frac{1}{H})$. Actually, when computing the indirect inference estimator

(50), one may be reluctant to use a very large H since it involves, for each value of θ within a minimization algorithm, computing H instrumental estimators $\beta_T^h(\theta), h = 1, ..., H$. In the next subsection, we introduce several techniques for replacing these H computations by only one. However, this comes at the price which is the likely loss of the nice finite sample properties of (50) and (51).

In conclusion, let us stress that indirect inference is able, beyond finite sample biases, to correct for any kind of misspecification bias. The philosophy of this method is basically to estimate a simple model, possibly wrong, to get easily an instrumental estimator $\widetilde{\beta}_T$ while a direct estimation of structural parameters θ would have been a daunting task. Therefore, what really matters is to use an instrumental parameters vector β which captures the key features of the parameters of interest θ, while being much simpler to estimate. For instance, Pastorello, Renault and Touzi (2000) as well as Engle and Lee (1996) have proposed to first estimate a GARCH model as an instrumental model to indirectly recover an estimator of the structural model of interest, a stochastic volatility model much more difficult to estimate directly due to the presence of latent variables and possibly a continuous time specification. Other natural examples are models with latent variables such that an observed variable provides a convenient proxy. An estimator based on this proxy suffers from a misspecification bias but we end up with a consistent estimator by applying the indirect inference matching. For instance, Pastorello, Renault and Touzi (2000) use Black and Scholes implied volatilities as a proxy of realizations of the latent spot volatility process.

5.3 Simulated method of moments

Simulated method of moments (SMM), as introduced by Ingram and Lee (1991) and Duffie and Singleton (1993), is the simulation-based counterpart of GMM designed to take advantage of the informational content of given moment restrictions:

$$E\{K[y_t, z(1,t)] \mid z(1,t)\} = k[z(1,t), \theta^0]$$

where $z(1,t)$ stands for a vector of predetermined variables. The role of simulations in this context is to provide a Monte Carlo assessment of the population conditional moment function $k[z(1,t), \theta]$ when it is not easily available in closed form. Thus, the natural way to extend GMM via a Monte Carlo evaluation of the population moment is to minimize, over the unknown parameter vector θ, a norm of the sample mean of:

$$Z_t\{K[y_t, z(1,t)] - \frac{1}{H}\sum_{h=1}^{H} K[y_t^h(\theta), z(1,t)]\}$$

where Z_t is a matrix of chosen instruments, that is a fixed matrix function of $z(1,t)$. It is then clear that the minimization program which is considered is a particular case of (50) above with:

$$\widetilde{\beta}_T = \frac{1}{T}\sum_{t=1}^{T} Z_t K[y_t, z(1,t)]$$

and $\beta_{T,H}(\theta)$ defined accordingly. In other words, we reinterpret SMM as a particular case of indirect inference, when the instrumental parameters to match are simple moments rather than themselves defined through some structural interpretations. Note however that the moment conditions for SMM could be slightly more general since the function $K[y_t, z(1,t)]$ itself could depend on the unknown parameters θ. In any case, the general asymptotic theory sketched above for SII is still valid.

By contrast with general SII as presented above, an advantage of SMM is that the instrumental parameters to match, as simple moments, are in general easier to compute than estimated auxiliary parameters $\beta_T^h(\theta), h = 1, ..., H$, derived from some computationally demanding extremum estimation procedure. Gallant and Tauchen (1996) have taken advantage of this remark to propose a practical computational strategy for implementing indirect inference when the estimator $\widetilde{\beta}_T$ of the instrumental parameters is obtained as a M-estimator solution of:

$$\max_{\beta} \frac{1}{T}\sum_{t=1}^{T} q_t[y(0,t), z(1,t), \beta]\,.$$

The key idea is then to define the moments to match through the (pseudo)-score vector of this M-estimator. Let us denote:

$$K[y(0,t), z(1,t), \beta] = \frac{\partial q_t}{\partial \beta}[y(0,t), z(1,t), \beta] \tag{52}$$

and consider a SMM estimator of θ obtained as a minimizer of the norm of a sample mean of:

$$K[y(0,t), z(1,t), \widetilde{\beta}_T] - \frac{1}{H}\sum_{h=1}^{H} K[y^h(0,t)(\theta), z(1,t), \widetilde{\beta}_T]\,.$$

For a suitable GMM metric, such a minimization defines a so-called simulated-score matching estimator $\widetilde{\theta}_T$ of θ. In the spirit of Gallant and Tauchen (1996), the objective function q_t which defines the initial estimator $\widetilde{\beta}_T$ typically is the log-likelihood of some auxiliary model. However, this feature is not needed for the validity of the asymptotic theory sketched below. Several remarks are in order.

(i) By contrast with a general SMM criterion, the minimization above does not involve the choice of any instrumental variable. Typically, over-identification will be achieved by choosing an auxiliary model with a large number of instrumental parameters β rather than by choosing instruments.

(ii) By definition of $\widetilde{\beta}_T$, the sample mean of $K[y(0,t), z(1,t), \beta]$ defined by (52) takes the value zero for $\beta = \widetilde{\beta}_T$. In other words, the minimization program above amounts to:

$$\min_{\theta} \| \frac{1}{TH} \sum_{t=1}^{T} \sum_{h=1}^{H} \frac{\partial q_t}{\partial \beta} [y^h(0,t)(\theta), z(1,t), \widetilde{\beta}_T] \|_{\Omega_T} \tag{53}$$

where the notation $\| \cdot \|_{\Omega_T}$ stands for a norm computed with a suitable GMM metric Ω_T.

(iii) It can be shown (see Gouriéroux, Monfort and Renault (1993)) that under the same assumptions as for the asymptotic theory of SII, the score matching estimator is consistent asymptotically normal. We get a score matching estimator with a minimum asymptotic variance if and only if the limit-weighting matrix Ω is proportional to the inverse of the asymptotic conditional variance of $\sqrt{T} \sum_{t=1}^{T} \frac{\partial q_t}{\partial \beta} [y(0,t), z(1,t), b(\theta^0)]$. Then the resulting efficient score matching estimator is asymptotically equivalent with the efficient indirect inference estimator.

(iv) Due to this asymptotic equivalence, the score-matching estimator can be seen as an alternative to the efficient SII estimator characterized in the previous subsection. This alternative is often referred to as Efficient Method of Moments (EMM) since, when $q_t[y(0,t), z(1,t), b(\cdot)]$ is the log-likelihood of some auxiliary model, the estimator is as efficient as maximum likelihood under correct specification of the auxiliary model. More generally, the auxiliary model is designed to approximate the true data generating process as closely as possible and Gallant and Tauchen (1996) propose the Semi-Non-Parametric (SNP) modeling to this end. These considerations and the terminology EMM should not lead to believe that Score-Matching is more efficient than Parameter-Matching Indirect Inference. The two estimators are asymptotically equivalent even though the Score Matching approach makes more transparent the required spanning property of the auxiliary model to reach the Cramér Rao efficiency bound of the structural model. Starting with the seminal papers of Gallant, Hsieh and Tauchen (1997) in discrete time and Andersen and Lund (1997) in continuous time, a large literature of estimation of stochastic volatilty models by EMM has been developed within the last ten years. Chernov and Ghysels (2000) have been one step further by using simultaneously data on the return process with stochastic volatility and on prices of options contracts written on this return.

(v) Another alleged advantage of score matching with respect to parameter matching in SII is its low computational cost. The fact is that with a large number of instrumental parameters β, as it will typically be the case with a SNP auxiliary model, it may be costly to maximize H times the log-likelihood of the auxiliary model (for each value of θ along an optimization algorithm) with respect to β to compute $\beta_T^h(\theta), h = 1, ..., H$. By contrast, the program (53) minimizes only once the norm of a vector of derivatives with respect to β. However it is worth realizing that not only is this cheaper computation likely to destroy the expected nice finite sample properties of SII put forward in the previous subsection, but also that the point is not really about a choice between matching (instrumental) parameters β or matching the (instrumental) score $\sum_{t=1}^{T} \frac{\partial q_t}{\partial \beta}$. The key issue is rather how to use H simulated paths, each of length T, as explained below.

(vi) The sum of TH terms considered in the definition (53) of the score-matching estimator is akin to considering only one simulated path $y^1(0, TH)(\theta)$ of size TH built from random draws as above. From such a simulated path, estimation of instrumental parameters would have produced a vector $\beta_{TH}^1(\theta)$ that could have been used for indirect inference, that is to define an estimator $\widetilde{\theta}_T$ solution of:

$$\min_{\theta} [\widetilde{\beta}_T - \beta_{TH}^1(\theta)]' \Omega_T [\widetilde{\beta}_T - \beta_{TH}^1(\theta)] \,. \tag{54}$$

This parameter matching estimator is not more computationally demanding than the corresponding score matching estimator computed from the same simulated path as solution of:

$$\min_{\theta} \parallel \frac{1}{TH} \sum_{t=1}^{TH} \frac{\partial q_t}{\partial \beta} [y^h(0,t)(\theta), z(1,t), \widetilde{\beta}_T] \parallel_{\Omega_T} \tag{55}$$

Actually, they are even numerically identical in the case of just-identification ($\dim \beta = \dim \theta$). More generally, the four estimators (50), (54), (53), (55) are asymptotically equivalent when T goes to infinity and the GMM weighting matrix are efficiently chosen accordingly. However, it is quite obvious that only (50) performs the right finite sample bias correction by matching instrumental parameters values estimated on both observed and simulated paths of lengths T. The trade off is thus between giving up finite sample bias correction or paying the price for computing H estimated instrumental parameters.

5.4 Indirect inference in presence of misspecification

The econometrician's search for a well-specified parametric model ("quest for the Holy Grail" as stated by Monfort (1996)) and associated efficient estimators even remain popular when maximum likelihood estimation becomes intractable due to highly non-linear structure including latent variables as typically the stochastic volatility models. The efficiency property of EMM and more generally of SMM and SII when the set of instrumental parameters to match is sufficiently large to span the likelihood score is often advocated as if the likelihood score was well specified. However, the likely misspecification of the structural model requires a generalization of the theory of SII as recently proposed by Dridi, Guay and Renault (2007). As for maximum likelihood with misspecification (see White (1982), Gouriéroux, Monfort and Trognon (1984)), such a generalization entails two elements.

First, asymptotic variance formulas are complicated by the introduction of sandwich formulas. Ignoring this kind of correction is even more detrimental than for QMLE since two types of sandwich formulas must be taken into account, one for the data generating process (DGP) and one for the simulator based on the structural model, which turns out to be different from the DGP in case of misspecification.

Secondly, and even more importantly, misspecification may imply that we consistently estimate a pseudo-true value, which is poorly related to the true unknown value of the parameters of interest. Dridi, Guay and Renault (2007) put forward the necessary (partial) encompassing property of the instrumental model (through instrumental parameters β) by the structural model (with parameters θ) needed to ensure consistency toward true values of (part of) the components of the estimated θ in spite of misspecification. The difficult issue is that, since structural parameters are recovered from instrumental ones by inverting a binding function $\beta = b(\theta)$, all the components are in general interdependent. The requirement of encompassing typically means that, if one does no want to proceed under the maintained assumption that the structural model is true, one must be parsimonious with respect to the number of moments to match or more generally to the scope of empirical evidence that is captured by the instrumental parameters β.

For instance, in an unpublished working paper, Dridi and Renault (see also Dridi's PhD thesis, University of Toulouse) show that for volatility leverage effects incorrectly modeled as return skewness or vice versa, only well focused instrumental parameters enable consistent estimation of the volatility persistence while supposedly efficient moment matching like EMM based on a SNP score generator would provide inconsistent estimators. Generally speaking, robustness to misspecification requires an instrumental model choice strategy quite opposite to the one commonly used for a structural model: the larger the instrumental model, the larger the risk of contamination of the estimated structural parameters of interest by what is wrong in the structural model. Of course there is no such thing as a free lunch: robustness to misspec-

ification through a parsimonious and well-focused instrumental model comes at the price of efficiency loss. Efficiency loss means not only lack of asymptotic accuracy of structural parameters estimates but also lack of power of specification tests. By contrast, an important advantage of the SNP score matching is to provide a battery of specification tests which, unfortunately, more often than not will lead to the conclusion that the structural model performs poorly in some directions and thus the allegedly efficient estimator must be viewed with a healthy dose of skepticism.

6 Concluding Remarks

Twenty years ago, the GARCH model and its many variants became temporarily dominant in the econometrics of financial time series. In these models, the conditional volatility is perfectly fitted by past observations, whereas the SV models allow for additional uncertainty in the volatility. As observed by Shephard (2005) about GARCH modelling, "this one-step ahead prediction approach to volatility modeling is very powerful, particularly in the field of risk management". Moreover, "it is convenient from an econometric viewpoint as it immediately delivers the likelihood function as the product of one-step ahead predictive densities". By contrast, the SV model is generally considered difficult to estimate even though it has regained some popularity recently because of the development of computationally intensive estimation methods, especially MCMC. In contrast, this chapter has reviewed alternative moment-based inference approaches to SV models. Besides their simplicity, the MM methods have the advantage of focusing on conditional moment restrictions which are excessively strict within a GARCH setting but are conveniently relaxed by adopting the SV framework.

Inspection of the burgeoning literature on SV leads us to conjecture that this chapter will need many complements in the next few years. As often in a survey paper, we have overlooked recent developments that sound promising. Let us sketch five important issues that are absent in this survey.

(i) The main reason why, in spite of involved technical issues, the SV approach has not been fully subsumed by GARCH is its usefulness for option pricing. There are at least two deep reasons for that. First, mainstream option pricing model are written in continuous time and SV models can naturally fit within the continuous time setting while the one-step ahead approach of GARCH is at odds with changes in the sampling frequency. Second, option markets are not redundant due to incompleteness of the underlying asset market. Latent volatility factors are an excellent way to accommodate such incompleteness. I acknowledge that econometrics of option pricing is largely absent from the present survey. One of the reasons for that is the current relatively minor role of moment based approaches in option pricing. The fact is that, while a key feature of MM is to be semi-parametric, option

pricing appears parametric in nature since one needs a risk neutral probability measure to price a multitude of derivative contracts. However, the recent new approach to GMM in terms of empirical likelihood may bridge the gap, as shown recently by Gagliardini, Gouriéroux and Renault (2007).

(ii) Another way to bridge the gap between option pricing and the MM approach to SV models is to characterize option prices as functions of moments of the appropriately defined notion of realized volatility. The presence of jumps and leverage effects may complicate the matter (see Ji, Renault and Yoon (2008)). Moreover, as mentioned in Section 4.2 of this survey, moment-based estimation with realized volatility is still in its infancy. We clearly need further work, for instance with the double asymptotic methodology of Corradi and Distaso (2006), to be able to write conditional moment restrictions about the appropriate notion of realized volatility in the presence of multiple latent volatility factors, leverage effects and jumps.

(iii) The aforementioned discussion about the appropriate notion of realized volatility paves the way for an even more general discussion about the suitable aggregators of high frequency data. Even though this issue has been briefly discussed in Section 4 of this survey, much more is needed. Besides the mentioned intraday patterns of seasonality in volatility, one should also take advantage of high frequency data to capture the possibly asymmetric responses to news (Andersen, Bollerslev, Diebold and Vega (2003,2007), Chen and Ghysels (2008)), the informational content of random dates of quote changes (Renault and Werker (2008)), the difference between Lévy-type jumps and sequences of jumps in crash periods, etc.

(iv) As acknowledged in the introduction, one should go even further by taking into account all the explanatory power for the volatility of returns of all observed variables other than the past realizations of returns themselves. Note that this is another advantage of the SV approach relative to GARCH. Since the focus is on the conditional variance given possibly latent conditioning information, it paves the way for incorporating any additional source of observed relevant information that may help to filter the volatility measure of interest.

(v) An even simpler way to increase the relevant conditioning information is to consider a set of joint asset returns, that is to write down a multivariate SV model. While multivariate issues are essentially absent in the present survey, the MM inference approach to SARV models put forward here is readily extended to a multivariate setting (see Doz and Renault (2006)). This is natural as the demands of parsimony suggest capturing volatility via common factors that are SV, since they are latent, even if they may be modeled as GARCH (see Diebold and Nerlove (1989)). Second, as for the literature on Factor Analysis, a key issue for multivariate volatility modeling with common factors is identification. It is then important to characterize the set of conditional moment restrictions that are needed for identification. This semi-parametric approach to identification is more satisfactory for our understanding of volatility dynamics than a fully parametric one. The point

is that a likelihood approach may spuriously identify the volatility factors through tight constraints among higher order moments enforced by the likelihood function but may have nothing to do with the actual volatility process.

References

Aguilar, M. and Renault, E. (2008): A Moment-based Test for GARCH against Stochastic Volatility. *Working Paper UNC*.

Altonji, J.G. and Segal, L.M. (1996): Small Sample Bias in GMM Estimation of Covariance Structures. *Journal of Business and Economic Statistics* **14**, 353–366.

Andersen, T. G. (1994): Stochastic Autoregressive Volatility: A Framework for Volatility Modeling. *Mathematical Finance* **4**, 75–102.

Andersen, T.G. and Bollerslev, T. (1997): Intraday Periodicity and Volatility Persistence in Financial Markets. *Journal of Empirical Finance* **4**, 115–158.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2006): Parametric and Nonparametric Volatility Measurement. *In: Aït-Sahalia, Y. and Hansen, L.P. (Eds.): Handbook of Financial Econometrics*. Elsevier Science B.V., Amsterdam forthcoming.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2007): Roughing It Up: Including Jump Components in Measuring, Modeling and Forecasting Asset Return Volatility. *Review of Economics and Statistics* **89**, 701–720.

Andersen, T., Bollerslev, T., Diebold, F.X. and Ebens, H. (2001): The distribution of stock return volatility. *Journal of Financial Economics* **61**, 43–76.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2001): The distribution of exchange rate volatility. *Journal of American Statistical Association* **96**, 42–55.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003): Modeling and forecasting realized volatility. *Econometrica* **71**, 579–626.

Andersen T.G., Bollerslev T., Diebold F.X. and Vega C. (2003): Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. *The American Economic Review* **93**, 38-62.

Andersen T.G., Bollerslev T., Diebold F.X. and Vega C. (2007): Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics* **73**, 251-277.

Andersen T.G., Bollerslev T„ and Meddahi N. (2005): Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. *Econometrica* **73**, 279-296

Andersen, T.G. and Lund, J. (1997): Estimating continuous–time stochastic volatility models. *Journal of Econometrics* **77**, 343–379.

Andersen, T.G. and Sorensen, B.E. (1996): GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Business and Economic Statistics* **14**, 328–352.

Andreou, E. and Ghysels, E. (2002): Rolling–Sample Volatility Estimators: Some New Theoretical Simulation and Empirical Results. *Journal of Business and Economic Statistics* **20**, 363–376.

Andrews, D. (1993): Exactly median unbiased estimation of first order autoregressive/unit root models. *Econometrica* **61**, 139–65.

Antoine B., Bonnal H. and Renault E. (2007): On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood. *Journal of Econometrics* **138**, 461-487.

Barndorff–Nielsen, O.E and Shephard, N. (2001): Non–Gaussian Ornstein–Uhlenbeck–based models and some of their uses in financial economics. *Journal of the Royal Statistical Society Series B* **63**, 167–241.

Barndorff–Nielsen, O.E. and Shephard, N. (2002): Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B* **64**, 253–280.

Barndorff–Nielsen, O.E and Shephard, N. (2004): Power and bipower variation with stochastic jumps. *Journal of Financial Econometrics* **2**, 1–48.

Barndorff–Nielsen, O.E. and Shephard, N. (2007): Variations, jumps, market frictions and high frequency data in financial econometrics. *In: Blundell, R., Torsten, P. and Newey, W.K. (Eds.): Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*. Econometric Society Monographs, Cambridge University Press.

Black, F. (1976): The pricing of commodity contracts. *Journal of Financial Economics* **3**, 167–79.

Black, F. and Scholes, M. (1973): The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637–659.

Bollerslev T„(1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T., Gibson, M. and Zhou, H. (2004): Dynamic estimation of volatility risk premia and investor risk aversion from option–implied and realized volatilities. *Finance and Economics Discussion Series, Divisions of Research and Statistics and Monetary Affairs Federal Reserve Board, Washington, D.C.*.

Bollerslev, T., Litvinova, J. and Tauchen, G. (2006): Leverage and Volatility Feedback Effects in High–Frequency Data. *Journal of Financial Econometrics* **4**, 353–384.

Bollerslev, T. and Zhou, H. (2002): Estimating stochastic volatility diffusion using conditional moments of the integrated volatility. *Journal of Econometrics* **109**, 33–65.

Carnero, M.A, Pena, D. and Ruiz, E. (2004): Persistence and Kurtosis in GARCH and Stochastic Volatility Models. *Journal of Financial Econometrics* **2**, 319–342.

Carrasco, M. and Florens, J.P. (2002): Simulation based method of moments and efficiency. *Journal of Business and Economic Statistics* **20**, 482–492.

Carrasco M, Chernov M., Florens J.P. and Ghysels E. (2007): Efficient Estimation of Jump Diffusions and General Dynamic Models with a Continuum of Moment Conditions. *Journal of Econometrics* **140**, 529–573.

Chen, X. and Ghysels, E. (2008): News - Good or Bad - and its impact on volatility predictions over multiple horizons. *Discussion Paper, UNC.*

Chernov, M. and Ghysels, E. (2000): A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of Financial Economics* **56**, 407–458.

Clark, P.K. (1973): A subordinated stochastic process with fixed variance for speculative prices. *Econometrica* **41**, 135–156.

Corradi, V. and Distaso, W. (2006): Semiparametric comparison of stochastic volatility models using realized measures. *Review of Economic Studies* **73**, 635–667.

Cox, D., Ingersoll, J.E. and Ross, S. (1985): A theory of the Term Structure of Interest Rates. *Econometrica* **53**, 385–407.

Diebold, F.X. and Nerlove, M. (1989): The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. *Journal of Applied Econometrics* **4**, 1–22.

Doz, C. and Renault, E. (2006): Factor Stochastic Volatility in Mean Models: a GMM approach. *Econometric Reviews* **25**, 275–309.

Dridi, R., Guay, A. and Renault, E. (2007): Indirect Inference and Calibration of Dynamic Stochastic General Equilibrium Models. *Journal of Econometrics* **136**, 397–430.

Drost, F.C. and Nijman, T.E. (1993): Temporal aggregation of GARCH processes. *Econometrica* **61**, 909–927.

Drost, F. and Werker, B. (1996): Closing the GARCH gap: continuous time GARCH modeling. *Journal of Econometrics* **74**, 31–58.

Duffie, D., Pan, J. and Singleton, K. (2000): Transform Analysis and Asset Pricing for Affine Jump-Diffusions. *Econometrica* **68**, 1342–1376.

Duffie, D. and Singleton, K. (1993): Simulated moments estimation of Markov models of asset prices. *Econometrica* **61**, 929–952.

Engle, R.F. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987–1008.

Engle, R.F. (1995): *ARCH: Selected Readings*. Oxford University Press.

Engle, R.F. (2000): The Econometrics of Ultra-High Frequency Data. *Econometrica* **68**, 1–22.

Engle, R. F., Ghysels, E. and Sohn, B. (2006): On the Economic Sources of Volatility. *Discussion Paper NYU and UNC*.

Engle, R. F. and Lee, G. G. J. (1999): A Permanent and Transitory Component Model of Stock Retrun Volatility. *In: Engle, R. F. and White, H. (Eds.): Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, 475–497. Oxford University Press.

Engle, R.F., Lilien, D.M and Robins, R.P. (1987): Estimating Time-Varying Risk Premia in the Term Structure: the ARCH-M Model. *Econometrica* **55**, 391–407.

Forsberg, L. and Ghysels, E. (2007): Why Do Absolute Returns Predict Volatility So Well. *Journal of Financial Econometrics* **5**, 31–67.

Francq, C. and Zakoïan, J.M. (2000): Estimating weak GARCH representations. *Econometric Theory* **16**, 692–728.

Francq, C. and Zakoïan, J.M. (2004): Maximum Likelihood Estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* **10**, 605–637.

Francq, C. and Zakoïan, J.M. (2006): Linear-Representation Based Estimation of Stochastic Volatility Models. *Scandinavian Journal of Statistics* **33**, 785–806.

Franses, P.H., Van der Leij, M. and Paap, R. (2008): A Simple Test for GARCH against a Stochastic Volatility Model. *Journal of Financial Econometrics* **6**, 291–306

French, K.F, Schwert, G.W. and Stambaugh, R.F. (1987): Expected Stock Returns and Volatility. *Journal of Financial Economics* **19**, 3–29.

Gagliardini, P., Gouriéroux, C. and Renault, E. (2007): Efficient Derivative Pricing by Extended Method of Moments. *Working Paper UNC*.

Gallant, A.R., Hsieh, D. and Tauchen, G. (1997): Estimation of Stochastic volatility Models with Diagnostics. *Journal of Econometrics* **81**, 159–192.

Gallant, A.R. and Tauchen, G. (1996): Which Moments to Match. *Econometric Theory* **12**, 657–681.

Garcia, R., Lewis, M.A., Pastorello, S. and Renault, E. (2007): Estimation of Objective and Risk Neutral Distributions based on Moments of Integrated Volatility. *Journal of Econometrics*, forthcoming.

Ghysels, E., Harvey, A. and Renault, E. (1996): Stochastic Volatility. *In: Maddala, G. S. and Rao, C. R. (Eds.): Handbook of Statistics* **14**. Elsevier Science B.V.

Ghysels,E., and Jasiak J. (1998): GARCH for irregularly spaced financial data: The ACD-GARCH model. *Studies in Nonlinear Dynamics and Econometrics* **2**, 133–149.

Ghysels, E., Mykland, P. and Renault, E. (2007): In-Sample Asymptotics and Across-Sample Efficiency Gains for High Frequency Data Statistics. *Working Paper*.

Ghysels, E., Santa–Clara, P. and Valkanov, R. (2005): There is a Risk–Return Tradeoff After All. *Journal of Financial Economics* **76**, 509–548.

Ghysels, E., Santa–Clara, P. and Valkanov, R. (2006): Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies. *Journal of Econometrics* **131**, 59–95.

Ghysels, E. and Sinko, A. (2006): Comments on Hansen and Lunde. *Journal of Business and Economic Statistics* **24**, 192–194.

Ghysels, E., Sinko, A. and Valkanov, R. (2006): MIDAS Regressions: Further Results and New Directions. *Econometric Reviews* **26**, 53–90.

Gourieroux, C. and Jasiak, J. (2006): Autoregressive Gamma Processes. *Journal of Forecasting* **25**, 129–152.

Gouriéroux, C. and Monfort, A. (1992): Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* **11**, 195–228.

Gouriéroux, C. and Monfort, A. (1996): Simulation–based econometric methods. *Core Lectures*, Oxford University Press.

Gouriéroux C., Monfort, A. and Renault, E. (1993): Indirect Inference. *Journal of Applied Econometrics* **8**, S85–S118.

Gouriéroux, C., Monfort, A. and Trognon, A. (1984): Pseudo–maximum likelihood methods theory. *Econometrica* **52**, 681–700.

Gouriéroux, C., Renault, E. and Touzi, N. (2000): Calibration by simulation for small sample bias correction. *In: Mariano, R., Schuerman, T. and Weeks, M.J. (Eds.): Simulation–Based Inference in Econometrics*. Cambridge University Press.

Granger, C.W.J. and Newbold, P. (1976): Forecasting transformed series. *Journal of the Royal Statistical Society Series B* **38**, 189–203.

Hall A.R. (2005): *Generalized Method of Moments* Oxford University Press, Oxford.

Hansen, L.P., Heaton, J.C. and Ogaki, M. (1988): Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions. *Journal of the American Statistical Association* **83**, 863–871.

Hansen, L.P., Heaton, J. and Yaron, A. (1996): Finite–Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics* **14**, 262–280.

Heston, S.L., (1993): A closed–form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies* **6**, 327–343.

Ingram, B.F. and Lee, B.S. (1991): Estimation by simulation of time series models. *Journal of Econometrics* **47**, 197–207.

Jacod J. (1994): Limit of random measures associated with the increments of a Brownian semimartingale. *Preprint number 120, Laboratoire de probabilites* Université Pierre et Marie Curie, Paris.

Jacod J. (1997): On continuous conditional Gaussian martingales and stable convergence in law. *In: Seminaire Probability, Lecture Notes in Mathematics* **1655**, 232–246. Springer Verlag.

Jacod J. and Shirayev, A.N. (1987): *Limit Theorems for Stochastic Processes*. Springer Verlag.

Jacquier E., Polson, N.G. and Rossi, P.E. (1994): Bayesian analysis of stochastic volatility models (with discussion): *Journal of Business and Economic Statistics* **12**, 371–417.

Ji, C., Renault, E. and Yoon, J. (2008): An Approximation Scheme for Option Pricing with Stochastic Volatility and Jumps. *Working Paper*.

Jiang, G. and Tian, Y. (2005): Model–free implied volatility and its information content. *Review of Financial Studies* **18**, 1305–1342.

Kim, S., Shephard, N. and Chib, S. (1998): Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies* **45**, 361–393.

Kobayashi, M. and Shi, X. (2005): Testing for EGARCH against Stochastic Volatility Models. *Journal of Time Series Analysis* **26**, 135–150.

Meddahi, N. (2001): An eigenfunction approach for volatility modeling. *Working Paper 29–2001, CRDE, Université de Montréal*.

Meddahi, N. (2002): Moments of continuous time stochastic volatility models. *Working Paper, Université de Montréal*.

Meddahi, N. and Renault, E. (2004): Temporal Aggregation of Volatility Models. *Journal of Econometrics* **119**, 355–379.

Meddahi, N., Renault, E. and Werker, B. (2006): GARCH and Irregularly Spaced Data. *Economic Letters* **90**, 200–204.

Melino, A. and Turnbull, S.M. (1990): Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* **45**, 239–265.

Monfort, A. (1996): A reappraisal of misspecified econometric models. *Econometric Theory* **12**, 597–619.

Nelson, D.B. (1991): Conditional heteroskedasticity in asset pricing: a new approach. *Econometrica* **59**, 347–370.

Newey, W.K. and Smith, R.J. (2004): Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica*, **72**, 219–255.

Pastorello, S., Renault, E. and Touzi, N. (2000): Statistical inference for random–variance option pricing. *Journal of Business and Economic Statistics* **18**, 358–367.

Renault, E. and Werker, B. (2008): Causality Effects in Return Volatility Measures with Random Times. *Working Paper*.

Rosenberg, B. (1972): The behaviour of random variables with nonstationary variance and the distribution of security prices. *WP11 GSBA, University of California, Berkeley*.

Ruiz, E. (1994): Quasi-Maximum Likelihood Estimation of Stochastic Volatility Models. *Journal of Econometrics* **63**, 284–306.

Schwert, G.W. (1989): Why Does Stock Market Volatility Change Over Time? *Journal of Finance* **44**, 1207–1239.

Shephard, N. (2005): *Stochastic Volatility: Selected Readings*. Oxford University Press, Oxford.

Smith, A.A. (1993): Estimating nonlinear time series models using simulated autoregressions. *Journal of Applied Econometrics* **8**, S63–S84.

Taylor, S. J. (1986): *Modelling Financial Time Series*. Wiley, Chichester.

White, H. (1982): Maximum likelihood estimation of mis–specified models. *Econometrica* **50**, 1–25.

Parameter Estimation and Practical Aspects of Modeling Stochastic Volatility

Borus Jungbacker and Siem Jan Koopman

Abstract Estimating parameters in a stochastic volatility (SV) model is a challenging task and therefore much research is devoted in this area of estimation. This chapter presents an overview and a practical guide of the quasi-likelihood and the Monte Carlo likelihood methods of estimation. The concepts of the methods are straightforward and the implementation is based on Kalman filter, smoothing, simulation smoothing, mode calculation and Monte Carlo simulation. These methods are general, transparent and computationally fast; therefore, they provide a feasible way for the estimation of parameters in SV models. Various extensions of the SV model are considered and some details are provided for the effective implementation of the Monte Carlo methods. Some empirical illustrations are given to show that the methods can be successful in measuring the unobserved volatility in financial time series.

1 Introduction

Volatility models are concerned with the analysis of time-varying characteristics of the variance in financial return series. The daily closure log prices of an asset or an index evolve usually as a random walk or a process close to it; therefore, the relative price changes often behave as a white noise series. This empirical finding is consistent with economic theory. When markets operate efficiently, all current information is consolidated in the price. The current

Borus Jungbacker
Department of Econometrics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, e-mail: bjungbacker@feweb.vu.nl

Siem Jan Koopman
Department of Econometrics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, e-mail: s.j.koopman@feweb.vu.nl

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_13,

asset price is then also the best forecast for the future asset price. The dynamics in the mean of asset prices can therefore be treated as not existing. On the other hand, it is well established that the unobserved volatility in asset returns is subject to a dynamic process. For example, a well-known empirical feature in finance and economics is the temporary clustering of large shocks in a series of daily returns. The clustering of such shocks implies serial correlation in the squared returns.

Although various ad hoc treatments are available for analyzing time-varying features in volatility, a model-based approach has become the industry standard. The observation-driven class of models for volatility dynamics is the well-known generalized autoregressive conditional heteroscedasticity (GARCH) model as developed and popularized by Engle (1982) and Bollerslev (1986). Their applications in empirical finance literature are exhaustive and their extensions in the econometrics and statistics literature are widespread; see Zivot (2008). The parameter-driven counterpart of GARCH is the class of stochastic volatility (SV) models as formulated by Taylor (1986) and Harvey et al. (1994). In various empirical studies it has been shown that the SV model provides a basis for more accurate forecasts of volatility than those provided by GARCH models; see Koopman et al. (2005). Furthermore, SV models have a closer connection with financial economics theory. For example, in the option pricing literature, the asset price is usually modeled by a stochastic differential equation (SDE) such as

$$\begin{aligned} \mathrm{d}\log P(t) &= \mu(t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_1(t), \\ \mathrm{d}\log \sigma^2(t) &= \left\{\gamma + (\phi - 1)\log \sigma^2(t)\right\}\,\mathrm{d}t + \sigma_\eta\,\mathrm{d}B_2(t), \end{aligned} \tag{1}$$

where $P(t)$ is the asset price at time t, $\mu(t)$ is a drift term and $\sigma(t)$ is the volatility of the asset price at time t. The drift term $\mu(t)$ is likely to be small and is in practice often set to zero. The stochastic properties of the mean and variance of $P(t)$ are determined by the independent Brownian motions $B_i(t)$, for $i = 1, 2$, and the unknown fixed coefficients γ, ϕ and σ_η. The basic SV model is a discrete-time version of the SDE (1).

The estimation of parameters in discretized SV models is not standard since a closed-form expression for the likelihood function does not exist; therefore, different approaches have been considered for inference in SV models. Estimation can be based on approximations (quasi-maximum likelihood), numerical methods for evaluating the likelihood (numerical integration) or simulation methods. Much focus in the econometrics and statistics literature is on the use of Bayesian Markov chain Monte Carlo (MCMC) methods for inference, see, for example, Jacquier et al. (1994) and Kim et al. (1998). In this chapter we focus on the Monte Carlo method of evaluating the likelihood function of the SV model. In particular, we adopt the methods of Shephard and Pitt (1997) and Durbin and Koopman (1997) based on importance sampling. Although these methods are not applicable in all situations, they are computationally fast (in comparison with most other simulation methods)

and relatively easy to implement. So when importance sampling methods can be implemented successfully, they can be regarded as an effective estimation methodology for SV models. Other estimation methods that have generated interest are the method of moments developed by Andersen and Sorensen (1996), frequency-domain estimation considered by Breidt et al. (1998) and the likelihood approaches explored by Fridman and Harris (1998) and Brockwell (2007). A general overview of the SV literature is given by the collection of articles in the book of Shepard (2005).

Consider a time series of asset log-returns y_t that is assumed to have constant mean μ and a time-variance variance $\exp h_t$. The observations are typically sampled at daily intervals. The basic version of the discretized SV model for y_t is given by

$$\begin{aligned} y_t &= \mu + \exp(\tfrac{1}{2}h_t)\varepsilon_t, \;\; \varepsilon_t \sim \mathrm{NID}(0,1), \\ h_{t+1} &= \gamma + \phi h_t + \eta_t, \qquad \eta_t \sim \mathrm{NID}(0,\sigma_\eta^2), \end{aligned} \tag{2}$$

for $t = 1, \ldots, n$ and where ε_t and η_s are independent of each other at all time points $t, s = 1, \ldots, n$. This basic SV model is a nonlinear time series model since both h_t and ε_t in the multiplication $\exp(\frac{1}{2}h_t)\varepsilon_t$ are stochastic. It is usually assumed that the log-volatility process is stationary but persistent, that is, $0 < \phi < 1$ is typically larger than 0.8. The unconditional mean of the log-volatility process is $(1-\phi)^{-1}\gamma$ and can be interpreted as the long-term log variance of the asset return series y_t. The unconditional variance of the log-volatility process is $(1-\phi^2)^{-1}\sigma_\eta^2$ and is sometimes referred to as the "volatility of volatility." Furthermore, the stochastic time-varying variance of the log returns y_t conditional on h_t is given by $\sigma_t^2 = \mathrm{E}(y_t - \mu)^2 = \exp h_t$, for $t = 1, \ldots, n$, where h_t can be any stationary autoregressive process. The conditional log density $p(y_t|h_t)$ is then given by

$$\log p(y_t|h_t) = -\frac{1}{2}\log 2\pi - \frac{1}{2}h_t - \frac{1}{2}\exp(-h_t)(y_t-\mu)^2, \qquad t = 1, \ldots, n. \tag{3}$$

The unknown coefficients that need to be estimated are γ, ϕ and σ_η and are collected in the parameter vector ψ.

The SV model is a nonlinear and non-Gaussian time series model with an observation equation and a state equation. The observation equation describes the relationship between the observations and the latent factors, while the state equation provides a model description for the dynamic properties of the latent factors. Denote $y = (y_1, \ldots, y_n)'$, where y_t is the scalar observation at time t for $t = 1, \ldots, n$, and denote $\theta = (\theta_1, \ldots, \theta_n)'$, where θ_t is the so-called signal at time t that is only determined by the latent factors. For the SV model (2), we have simply $\theta_t = h_t$. Furthermore, the observation density is given by $p(y|\theta)$ and the signal density is given by $p(\theta)$. For the classes of SV models in this chapter, the conditional independence assumption applies to the density $p(y|\alpha)$, that is,

$$p(y|\theta) = \prod_{t=1}^{n} p(y_t|\theta_t).$$

The model densities $p(y|\theta)$ and $p(\theta)$ depend on a set of unknown coefficients. In the case of (2), these are γ, ϕ and σ_η. The estimation of these coefficients will be based on maximum likelihood. The likelihood function can be expressed as

$$p(y) = \int p(\theta, y)\,\mathrm{d}\theta = \int p(y|\theta)p(\theta)\,\mathrm{d}\theta. \tag{4}$$

which is an n-fold integral and is typically not tractable except in the most trivial cases. An analytical expression for $p(y)$ is therefore not available for the SV class of models. We need to rely on numerical techniques for the evaluation of $p(y)$. In this Chapter we consider the method of Monte Carlo integration based on importance sampling. A straightforward Monte Carlo estimator of $p(y)$ in (4) is $\widetilde{p}(y) = M^{-1}\sum_{m=1}^{M} p(y|\theta^m)$ where θ^m is a draw from $p(\theta)$ and with $\widetilde{p}(y) \to p(y)$ as $M \to \infty$. However, this Monte Carlo estimator is not efficient since many draws from $p(\theta)$ will make no contribution to $p(y|\theta)$ and hence the estimate will be poor even for extremely high values of M.

In this article we opt for two approaches that overcome the problems as described above. First we consider a linearization of the nonlinear observation equation. This will lead to a transformation of y_t that has a linear observation equation. Obviously, this approach is an approximation to the SV model. Second we consider the Monte Carlo evaluation of the likelihood function. The method of importance sampling is considered for the evaluation of (4). In both approaches, state-space methods for the linear Gaussian state-space model play a prominent role. We therefore consider a general state-space representation of the SV model although the implications of the results presented for the SV model will be given explicitly. The state-space algorithms are instrumental but do not need to be discussed in much detail since we only need to apply them. As a service to the reader, the algorithms are briefly discussed in the Appendix. Since the SV model is discussed within a more general setting, various extensions of the SV model can be considered as special cases too. The general method applies to each of them but there are some differences in detail which will be reported. To illustrate the effectiveness of the methods, some empirical illustrations are given for stock index return series and for exchange rate series.

2 A Quasi-Likelihood Analysis Based on Kalman Filter Methods

The basic SV model (2) is intrinsically a nonlinear model owing to the multiplication of two stochastic variables in the observation equation, that is,

$y_t - \mu = \exp(\frac{1}{2}h_t)\varepsilon_t$. Since the sample mean of the log returns y_t is a consistent estimator of μ, we can replace μ by the sample mean. Harvey et al. (1994) have pointed out that the basic SV model (2) can be analyzed on the basis of a linearized version of the model. For this purpose, we consider scalar x_t as the transformation of y_t and given by

$$x_t = \log(y_t - \bar{y})^2, \qquad \bar{y} = n^{-1}\sum_{t=1}^{n} y_t, \tag{5}$$

for $t = 1, \ldots, n$. Given the basic SV model for y_t, a reasonable suggestion of a model for x_t is given by

$$x_t = \kappa_1 + h_t + u_t, \qquad h_{t+1} = \gamma + \phi h_t + \eta_t, \tag{6}$$

where $u_t = \log \varepsilon_t^2 - \kappa_1$ is distributed by the centered $\log \chi^2$ density with one degree of freedom. The mean and variance of $\log \varepsilon_t^2$ are given by κ_1 and κ_2, respectively. In this case, $\kappa_1 \approx -1.27$ and $\kappa_2 = \pi^2 \,/\, 2$. Model (6) is linear and the observation disturbance has a non-Gaussian density.

However, we may consider u_t also as a sequence of independent noise terms with mean zero and variance κ_2 to avoid the need to treat the density function of u_t explicitly. Linear methods can then be applied to obtain estimators of h_t that belong to the class of minimum mean squares linear estimators. The metric for estimation can nevertheless be chosen as the Gaussian likelihood function. Such an approach will be referred to as a quasi-maximum likelihood analysis. It effectively considers model (6) with Gaussian disturbance terms for the transformed series x_t, that is,

$$\begin{aligned} x_t &= \kappa_1 + h_t + u_t, \qquad h_{t+1} \;=\; \gamma + \phi h_t + \eta_t, \\ u_t &\sim \mathrm{NID}(0, \kappa_2), \qquad \eta_t \;\sim\; \mathrm{NID}(0, \sigma_\eta^2), \end{aligned} \tag{7}$$

for $t = 1, \ldots, n$. This linearized SV model is an example of the linear Gaussian state space model with the state equation for h_t and the observation equation for x_t. Although the SV model usually has the log volatility h_t modeled as an autoregressive (AR) process (mean-reverting), the state-space framework allows it to be modeled by many other linear Gaussian time series processes.

In the state-space formulation, we define θ_t as the signal and α_t as the state vector. A general observation equation for x_t can be given by

$$x_t = \theta_t + u_t, \qquad u_t \sim \mathrm{NID}(0, H_t), \tag{8}$$

for $t = 1, \ldots, n$. The dynamic properties of the signal are modeled by

$$\theta_t = c_t + Z_t\alpha_t, \qquad \alpha_{t+1} = d_t + T_t\alpha_t + \eta_t, \qquad \eta_t \sim \mathrm{NID}(0, Q_t), \tag{9}$$

for $t = 1, \ldots, n$, where system vectors c_t and d_t and system matrices Z_t, H_t, T_t and Q_t are fixed and known functions of parameter vector ψ. The

observation x_t and signal θ_t are scalar variables, while the state vector α_t together with the disturbance vector η_t have dimensions $q \times 1$. It follows that model (7) is represented by $x_t = \theta_t + u_t$, where $\theta_t = h_t + \kappa_1$ is modeled by (9) with $c_t = \kappa_1$, $Z_t = 1$, $H_t = \kappa_2$, $\alpha_t = h_t$, $d_t = \gamma$, $T_t = \phi$ and $Q_t = \sigma_\eta^2$. The linearized SV model is therefore a time-invariant case of the general linear Gaussian state space model (8) and (9). The Kalman filter and related methods can be applied to this state-space model; see the Appendix.

Stationary autoregressive moving averages processes as well as nonstationary linear processes for the signal θ_t can be formulated as (9). The system vectors and matrices in (9) have appropriate dimensions and the variance matrix Q_t is positive-semidefinite. The initial state vector is normally distributed with mean a and variance matrix P. The disturbances η_t are serially independent and are independent of the initial state vector for $t = 1, \ldots, n$. The joint property of a sequence of n state vectors can be expressed by the multivariate normal density

$$\alpha \sim \mathrm{N}(d, \Omega), \tag{10}$$

where

$$\alpha = (\alpha_1', \ldots, \alpha_n')', d = T\left(a', d_1', \ldots, d_{n-1}'\right)', \Omega = T\mathrm{diag}(P_1, Q_1, \ldots, Q_{n-1})T',$$

with

$$T = \begin{bmatrix} I & 0 & 0 & \cdots & 0 & 0 \\ T_1 & I & 0 & \ldots & 0 & 0 \\ T_2T_1 & T_2 & I & & 0 & 0 \\ & & & \ddots & & \vdots \\ T_{n-2}\ldots T_1 & T_{n-2}\ldots T_2 & T_{n-2}\ldots T_3 & & I & 0 \\ T_{n-1}\ldots T_1 & T_{n-1}\ldots T_2 & T_{n-1}\ldots T_3 & \cdots & T_{n-1} & I \end{bmatrix}, \tag{11}$$

for $t = 1, \ldots, n$.

It further follows that $\theta = (\theta_1, \ldots, \theta_n)'$ has a multivariate normal distribution given by

$$\theta \sim \mathrm{N}(\mu, \Psi), \qquad \mu = c + Zd, \qquad \Psi = Z\Omega Z', \tag{12}$$

where

$$\theta = c + Z\alpha, \qquad c = (c_1', \ldots, c_n')', \qquad Z = \mathrm{diag}(Z_1, \ldots, Z_n).$$

The log density of the signal is given by

$$\log p(\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\theta - \mu)'\Psi^{-1}(\theta - \mu). \tag{13}$$

The prediction error decomposition can be used to evaluate (13). Most of the commonly used linear Gaussian time series models, including autoregres-

sive moving average models, can be represented in state-space form with a positive-definite variance matrix Ψ.

Define $n \times 1$ vector $x = (x_1, \ldots, x_n)'$, then

$$x = \theta + u, \qquad u \sim \mathrm{N}(0, H), \tag{14}$$

with $u = (u_1, \ldots, u_n)'$ and $H = \mathrm{diag}(H_1, \ldots, H_n)$. In case of (7), we have $H_t = \kappa_2$ for $t = 1, \ldots, n$ and $H = \kappa_2 I_n$. The linear Gaussian observation density is given by

$$p(x|\theta) = \mathrm{N}(\theta, H) = \prod_{t=1}^{n} p(x_t|\theta_t). \tag{15}$$

We have shown that the linearized SV model can be represented by the general linear Gaussian state space model with signal density $p(\theta)$ and observation density $p(x|\theta)$.

2.1 Kalman filter for prediction and likelihood evaluation

The linear Gaussian state-space model as formulated in (8) and (9) can be analyzed using computationally efficient and fast recursive algorithms. These methods can be applied to any linear time series model in which the dynamics can be represented in the Markovian form. The computations are carried out by so-called order-n operations, where n is the number of observations. The main attractions of the state-space approach are its generality, its effective treatment of missing observations, its handling of other messy features in time series analysis and its natural way of carrying out one-step-ahead prediction and long-term forecasting.

The Kalman filter is given in the Appendix. It evaluates the estimator of the state vector α_t conditional on the past observations $X_{t-1} = \{x_1, \ldots, x_{t-1}\}$ together with its conditional variance as given by

$$a_t = E(\alpha_t|X_{t-1}), \qquad P_t = \mathrm{Var}(\alpha_t|X_{t-1}) = E[(\alpha_t - a_t|X_{t-1})(\alpha_t - a_t|X_{t-1})'],$$

respectively, for $t = 1, \ldots, n$. It follows that the prediction of the signal is given by $c_t + Z_t a_t$ with mean square error $Z_t P_t Z_t'$. Further, the observation prediction error and its conditional variance are given by

$$v_t = x_t - E(x_t|X_{t-1}) = x_t - c_t - Z_t a_t, \qquad F_t = E(v_t v_t'|X_{t-1}) = Z_t P_t Z_t' + H_t,$$

respectively, for $t = 1, \ldots, n$. The prediction error decomposition allows the log-likelihood function $\ell(\psi)$ to be evaluated analytically as

$$\ell(\psi) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log F_t - \frac{1}{2}\sum_{t=1}^{n} v_t' F_t^{-1} v_t,$$

where ψ is the parameter vector containing the coefficients γ, ϕ and σ_η. By maximizing the log likelihood $\ell(\psi)$ with respect to ψ, we obtain estimates for the unknown coefficients. Analytical expressions for the estimates at the maximum of $\ell(\psi)$ are not available and therefore we rely on numerical optimization methods such as quasi-Newton methods; see Nocedal and Wright (1999). In the case where x is modeled by the Gaussian density $p(x|\theta)$ as in (14), the estimate of ψ is referred to as the maximum likelihood estimate. However, in the case of the linearized SV model (7) for x_t, the estimate of ψ, obtained by maximizing the Gaussian likelihood function, is referred to as the quasi-maximum likelihood estimate since the assumption of a $\log \chi^2$ density for u_t is replaced by the assumption of a Gaussian density.

Interest in the analysis of log returns often focuses on the forecasting of volatility for the purpose of, for example, determining option prices in the short-term future. Volatility forecasting is based on the prediction of the future signal conditional on the observed time series $X_n = \{x_1, \ldots, x_n\}$, together with its conditional variance, as defined by

$$\widehat{\theta}_{n+j|n} = E(\theta_{n+j}|X_n), \qquad V_{n+j|n} = Var(\theta_{n+j}|X_n), \qquad j = 1, 2, 3 \ldots.$$

The forecasts of the signal from a general state-space model can also be obtained by the Kalman filter. This is due to the ability of the Kalman filter to deal with missing observations whether they occur within sample or out of sample; see the Appendix.

2.2 Smoothing methods for the conditional mean, variance and mode

The actual measurement of the unobserved volatility given all observed log returns is referred to as smoothing or signal extraction in the state-space terminology. More specifically, we aim to evaluate the estimated log volatility as defined by the conditional mean of θ conditional on the observed log returns x (or X_n). These smoothed estimators can be obtained within the general framework of the state-space model as follows. Given parameter estimates of ψ, the unconditional mean, variance and covariance of observation x and signal θ are given by

$$E(x) = \mu, \qquad \mathrm{Var}(x) = \Sigma = \Psi + H, \qquad \mathrm{Cov}(\theta, x) = \Psi. \tag{16}$$

It follows from the standard lemma of the multivariate normal density that the conditional means and variances are given by

$$E(\theta|x) = \widehat{\theta} = \mu + \Psi\Sigma^{-1}(x-\mu), \qquad \mathrm{Var}(\theta|x) = V = \Psi - \Psi\Sigma^{-1}\Psi. \quad (17)$$

The Kalman filter and smoother evaluate the conditional mean $E(\theta_t|x)$ and variance $\mathrm{Var}(\theta_t|x)$ in a recursive and computationally efficient way for a linear Gaussian state-space model, see the Appendix. More specifically, the smoothing algorithm is a backwards-operating recursion that accommodates output of the Kalman filter and it evaluates the quantities

$$\widehat{\theta}_t = E(\theta_t|x), \qquad V_t = \mathrm{Var}(\theta_t|x), \qquad t = 1, \ldots, n.$$

The evaluation of V_t enables the construction of confidence intervals for $\widehat{\theta}_t$ for $t = 1, \ldots, n$.

Since all densities are Gaussian, the conditional or posterior mode of $p(\theta|x)$, denoted by $\widetilde{\theta}$, is equivalent to the conditional mean of $p(\theta|x)$, that is, $\widetilde{\theta} = E(\theta|x)$. After some minor manipulations, it follows from the first equation in (17) that

$$\widetilde{\theta} = \left(\Psi^{-1} + H^{-1}\right)^{-1}\left(H^{-1}x + \Psi^{-1}\mu\right). \quad (18)$$

It should be emphasized that the Kalman filter and smoother effectively computes $\widetilde{\theta}$ for the linear Gaussian state-space model.

2.3 Practical considerations for analyzing the linearized SV model

The linearized SV model (7) provides a full statistical analysis of log returns by using standard state-space methods. The embedding of the model within the general state-space framework also allows the treatment of messy features in the time series such as missing values, irregularly spaced data, outliers and breaks; see Harvey et al. (1998). Furthermore, the methods are computationally efficient and fast. The necessary computations can be carried out by software packages such as the user-friendly STAMP program of Koopman et al. (2007) and the `SsfPack` library for `Ox` and `S-PLUS`; see Koopman et al. (1999) and Zivot et al. (2004).

However, there is the practical inconvenience that the log returns are transformed by taking logs of the squares (mean-adjusted) as in (5). In the case where the return is very small or even zero (whether or not it is mean-adjusted), this transformation can clearly lead to numerical problems. The problem is often referred to as the inlier problem and has been investigated by Breidt and Carriquiry (1996). They found that a transformation (based on a Taylor series) suggested by Fuller (1996) has proved to be more stable. This transformation is given by

$$x_t = \log(y_t^2 + s) - \frac{s}{y_t^2 + s}, \qquad t = 1, \ldots, n,$$

where s is the sample variance of y_t scaled by a small multiple, say, 0.02. In practical work, this transformation of x_t is preferred over (5). In the approaches described below, we do not require a transformation since the methods work with the untransformed process; see also the discussion in Davis and Rodriguez-Yam (2005).

The mean κ_1 and variance κ_2 of $\log \varepsilon_t^2$ are considered as given in the quasi-maximum likelihood procedure. These moments have known values and do not need to be estimated. However, from Monte Carlo simulation studies it has emerged that the small-sample properties for the estimates of ψ improve when κ_1 and κ_2 are assumed unknown and are estimated together with the coefficients in ψ.

Despite the effectiveness of the linearized SV model, it is unsatisfactory in that the underlying assumptions of the SV model are not properly treated by the approximation. Therefore, other methods have been developed to analyze the SV model without linearization. The remaining part of this contribution presents and discusses methods for the treatment of the SV model based on the nonlinear formulation (2).

3 A Monte Carlo Likelihood Analysis

Given the nonlinear property of the SV model (2), the likelihood function of the SV model does not have a convenient and analytical expression in terms of parameter vector ψ. The likelihood function can be expressed generally by

$$\ell(\psi) = p(y; \psi) = \int_{-\infty}^{\infty} p(y, \theta; \psi)\, \mathrm{d}\theta, \tag{19}$$

where y and θ are vectors formed by stacking the observations and states y_t and θ_t, respectively, as defined earlier. Given the potentially high dimensional vectors y and θ, direct numerical evaluation of the integral is not feasible. We therefore consider Monte Carlo methods for the evaluation of likelihood function (19). In particular, the method of importance sampling is explored in detail. For this purpose we adopt the trivial identity

$$p(y; \psi) = \int p(y, \theta; \psi)\, \mathrm{d}\theta = \int \frac{p(y, \theta; \psi)}{f(\theta; y, \psi)} f(\theta; y, \psi)\, \mathrm{d}\theta = \mathrm{E}_f \left(\frac{p(y, \theta; \psi)}{f(\theta; y, \psi)} \right), \tag{20}$$

where $\mathrm{E}_f(\cdot)$ denotes expectation with respect to some proposal density $f(\theta; y, \psi)$ that is chosen to be as close as possible to $p(\theta|y; \psi)$ but has convenient properties. The expectation $\mathrm{E}_f(\cdot)$ can be evaluated by sampling from $f(\theta; y, \psi)$ and averaging the importance weights $p(y, \theta; \psi)$ / $f(\theta; y, \psi)$.

In the case where we choose $f(\theta; y, \psi) = p(\theta|y; \psi)$, expression (20) reduces obviously to $p(y; \psi)$, which cannot be evaluated analytically. Furthermore, exact sampling from $p(\theta|y; \psi)$ is usually not feasible in the context of non-Gaussian and nonlinear state-space models such as the SV model (2). For the importance sampling evaluation of (20), we need to find a proposal density $f(\theta; y, \psi)$ that is sufficiently close to $p(\theta|y; \psi)$ but from which it is relatively easy to simulate. In most practical applications of importance sampling, the proposal density is found within the class of linear Gaussian densities. In the context of state-space models, the proposal $f(\theta; y, \psi)$ is usually set equal to the multivariate normal $\mathrm{N}(\widetilde{\theta}, V)$, where $\widetilde{\theta}$ is some location measure of θ and V is an appropriate variance matrix. The notation deliberately suggests that location and scale are related to (18) and the second equation of (17), respectively, as will become apparent later. Sampling from this multivariate normal density can be carried out by the simulation smoother; see the Appendix. This method is fast and straightforward and it provides the means for the evaluation of the likelihood (19) by importance sampling.

The remainder of this section discusses the construction of an effective proposal density, sampling from this density and the practical details of evaluating the likelihood function (20) by importance sampling. Other aspects of analyzing an SV model are also discussed, including the measurement and forecasting of volatility.

3.1 Construction of a proposal density

We adopt a multivariate normal density as the proposal density with its mean equal to the mode of the smoothing density $p(\theta|y)$ and its curvature equal to that of $p(\theta|y)$ at the mode. This choice of the proposal density $f(\theta; y)$ is made since it may be sufficiently close to the smoothing density $p(\theta|y)$. It requires us to find the mode $\widehat{\theta}$, by maximizing the smoothing density $p(\theta|y)$ with respect to θ, and its Hessian matrix G, that is $f(\theta; y) = \mathrm{N}(\widehat{\theta}, V)$, where $V = -G^{-1}$. For the basic SV model (2) we consider the general nonlinear non-Gaussian observation model $p(y|\theta)$ and the linear Gaussian signal vector $p(\theta)$. For this class of models, an analytical expression for the posterior mode $\widehat{\theta}$ of $p(\theta|y)$ is not available. We therefore obtain the mode by maximizing $p(\theta|y)$ with respect to θ using the Newton–Raphson method of optimization; see Nocedal and Wright (1999) for a treatment of numerical optimization methods. The dimension of θ is typically $n \times 1$, so matrix dimensions can be high and straightforward matrix manipulations become infeasible; therefore efficient algorithms need to be considered.

For a given guess g of the solution for $\widehat{\theta}$, the Newton–Raphson method proposes the new guess as

$$g^+ = g - \left[\ddot{p}(\theta|y)|_{\theta=g}\right]^{-1} \dot{p}(\theta|y)|_{\theta=g}\,, \tag{21}$$

where we define

$$\dot{p}(\cdot|\cdot) = \frac{\partial \log p(\cdot|\cdot)}{\partial \theta}, \qquad \ddot{p}(\cdot|\cdot) = \frac{\partial^2 \log p(\cdot|\cdot)}{\partial \theta \partial \theta'}. \tag{22}$$

Since $\log p(\theta|y) = \log p(y|\theta) + \log p(\theta) - \log p(y)$, we have

$$\dot{p}(\theta|y) = \dot{p}(y|\theta) - \Psi^{-1}(\theta - \mu), \qquad \ddot{p}(\theta|y) = \ddot{p}(y|\theta) - \Psi^{-1}. \tag{23}$$

The conditional independence assumption of the observation model implies that $\ddot{p}(y|\theta)$ is a block diagonal matrix.

The Newton–Raphson updating step reduces to

$$\begin{aligned} g^+ &= g - \left[\ddot{p}(y|\theta)|_{\theta=g} - \Psi^{-1}\right]^{-1} \left(\dot{p}(y|\theta)|_{\theta=g} - \Psi^{-1}\{g - \mu\}\right) \\ &= \left[\Psi^{-1} - \ddot{p}(y|\theta)|_{\theta=g}\right]^{-1} \left(\dot{p}(y|\theta)|_{\theta=g} - \ddot{p}(y|\theta)|_{\theta=g}\, g + \Psi^{-1}\mu\right) \\ &= \left(\Psi^{-1} + A^{-1}\right)^{-1} \left(A^{-1}x + \Psi^{-1}\mu\right), \end{aligned} \tag{24}$$

where

$$A = -\left[\ddot{p}(y|\theta)|_{\theta=g}\right]^{-1}, \qquad x = g + A\,\dot{p}(y|\theta)|_{\theta=g}\,. \tag{25}$$

We note the similarity of (24) and (18). In the case where $\ddot{p}(y|\theta)$ is negative-semidefinite for all θ, it follows that the Kalman filter and smoother can be used to compute g^+ by applying it to a state-space model with observation equation (14) for x_t as in (25) and $H = A$. The computation of $E(\theta|x)$ for this model returns g^+ as a result. This approach was taken by Shephard and Pitt (1997), Durbin and Koopman (1997) and So (2003). The mode $\widehat{\theta}$ for a non-Gaussian nonlinear observation model is obtained by the Newton–Raphson method where for each step the Kalman filter and smoother computes the new guess g^+. The Hessian matrix of the mode estimator $\widehat{\theta}$ is given by $G = \ddot{p}(\theta|x) = -\Psi^{-1} - A^{-1}$.

This approach of finding the mode $\widehat{\theta}$ is clearly not valid when $p(x|\theta)$ is not log-concave, so $\ddot{p}(x|\theta)$ is not negative-definite. This implies that the variance matrix H for the observation model (14) is not positive-definite. However, it was argued by Jungbacker and Koopman (2007) that in cases where $\ddot{p}(y|\theta)$ is not negative-definite, the Kalman filter and smoothing recursions can still be used for the computation of (24). The special structure of variance matrix $\Psi = Z\Omega Z'$ of the signal θ nevertheless allows the use of decompositions based on triangular matrices such as T in (11). Hence, it can be shown that the Kalman and smoothing recursions can also be used for the general model.

Although matrix $\Psi + A$ can be indefinite, the Hessian matrix $-\Psi^{-1} - A^{-1}$ should always be seminegative-definite for θ at or in the close neighborhood of $\widehat{\theta}$ by construction. In cases where the Hessian matrix is not negative-definite,

the Newton–Raphson step does not progresses to the maximum of $p(\theta|y)$ with respect to θ. To enforce global convergence, the algorithm can be modified by line-search and other numerical methods; see Nocedal and Wright (1999). In general, line-search strategies often speed up the maximization and stabilize the algorithm. A line search can be implemented by introducing the scalar $0 < \lambda \leq 1$ in (21) and defining

$$g_\lambda^+ = g - \lambda \left[\ddot{p}(\theta|y)|_{\theta=g} \right]^{-1} \dot{p}(\theta|y)|_{\theta=g} \,. \tag{26}$$

The line search consists of finding a value for λ so that

$$p(\theta|y)|_{\theta=g_\lambda^+} > p(\theta|y)|_{\theta=g} \,.$$

By combining (26) and (21), the line search computations are straightforward and are given by

$$g_\lambda^+ = g + \lambda(g^+ - g),$$

where $g^+ = g_\lambda^+\big|_{\lambda=1}$ is computed by (24) only once for different values of $0 < \lambda \leq 1$. Global convergence is ensured when an appropriate set of regularity conditions for the line search is fulfilled; see Nocedal and Wright (1999) for a detailed discussion. To check these conditions, it is usually necessary to evaluate the score function.

The score vector of $p(\theta|y)$ is defined in (22) and is given by (23), that is,

$$\frac{\partial \log p(\theta|y)}{\partial \theta} = \dot{p}(\theta|y) = \dot{p}(y|\theta) - \Psi^{-1}(\theta - \mu).$$

The score can also be evaluated by the Kalman filter and smoother algorithms; see Jungbacker and Koopman (2007). Given the computational device for evaluating the score, other maximization methods may also be considered to obtain the mode of $p(\theta|y)$. It is noted that different numerical problems can occur during the maximization of $p(\theta|y)$ with respect to the high-dimensional vector θ. Although line-search methods can stabilize the Newton–Raphson method, it may be necessary to switch to other score-based or quasi-Newton optimization methods. Therefore, this computationally efficient method of computing the score is important in practical work.

3.2 Sampling from the importance density and Monte Carlo likelihood

The likelihood function $\ell(\psi) = p(y) = \int p(\theta, y)\, \mathrm{d}\theta$ is estimated via importance sampling using the expression

$$\widehat{\ell}(\psi) = M^{-1} \sum_{i=1}^{M} \frac{p(\theta^i, y)}{f(\theta^i; y)}, \qquad \theta^i \sim f(\theta^i; y). \tag{27}$$

The computation of (27) requires algorithms to sample from the importance density $f(\theta^i; y)$ and to evaluate for $i = 1, \ldots, M$ the so-called importance weight $p(\theta^i, y) \,/\, f(\theta^i; y)$. In cases where the observation density $p(y|\theta)$ is log-concave, the importance density can be represented as a linear Gaussian state space model and the simulation smoothers of de Jong and Shephard (1995) and Durbin and Koopman (2002) can be used to simulate from $f(\theta; y)$ in a computationally efficient way. The derivations of these methods rely on a properly defined linear Gaussian observation equation with positive definite matrices for Σ and $H = A$. In the general case, in particular when $p(y|\theta)$ is not log-concave, simulations can be carried out by the modified simulation smoother proposed by Jungbacker and Koopman (2007), see Appendix. We note that the Hessian of $p(\theta|y)$ is evaluated at $\theta = \widehat{\theta}$ and therefore matrix G is guaranteed to be negative definite and V is positive definite as a result.

The computation of (27) further requires evaluating the importance weight $p(\theta^i, y) \,/\, f(\theta^i; y)$ for $i = 1, \ldots, M$. Given the linear Gaussian signal vector θ, the evaluation of the nominator is based on the identity $p(\theta, y) = p(y|\theta)p(\theta)$, where $p(y|\theta)$ is defined by the model and is usually straightforward to compute. The density of the signal $p(\theta)$ for $\theta = \theta^i$ is evaluated by the Kalman filter since $\theta = c + Z\alpha$ has the Markov property and the prediction error decomposition can be applied to $p(\theta)$. Given the draw $\theta^i \sim f(\theta; y)$ with $f(\theta; y) = \mathrm{N}(\widehat{\theta}, V)$ obtained from the simulation smoother, the denominator $f(\theta^i; y)$ in (27) can be evaluated using the output of the simulation smoother; see the Appendix.

The estimator (27) is subject to Monte Carlo error. A strong law of large numbers insists that $\widehat{\ell}(\psi) \to \ell(\psi)$ as $M \to \infty$ with a rate of convergence depending on the precision of the proposal density; see Geweke (1989). The choice of M can be relatively small when an accurate proposal density is chosen. To identify appropriate proposal densities we may rely on a set of tests and diagnostics; see Koopman et al. (2007). In practical work, it is often sufficient to take M equal to 100 or 500.

For the purpose of parameter estimation, we maximize the Monte Carlo estimator of the likelihood function $\widehat{\ell}(\psi)$ in (27) with respect to the unknown vector ψ. The Newton–Raphson method can be used to maximize the likelihood function directly; see also the discussion in Section 2. During the numerical search, the same random seed is used each time the Monte Carlo likelihood function (27) is computed for a different ψ. This ensures that $\widehat{\ell}(\psi)$ is continuous in ψ.

4 Some Generalizations of SV Models

The basic SV model (2) was the motivating example for the discussion of estimation methods in the previous section. However, these methods can also be adopted for a wider set of SV models. Some generalizations of the SV model are presented in this section together with the details of implementing the estimation methods.

We consider the observed time series of asset log returns y_t which are typically sampled at daily intervals but not necessarily. A general SV model is given by

$$y_t = \mu_t + \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim IID(0,1), \qquad t = 1, \ldots, n, \tag{28}$$

where σ_t is the unobserved volatility process and μ_t is the possibly time varying mean of log returns. In this chapter we assume

$$\sigma_t = \sigma \exp\left(\frac{1}{2}\theta_t\right),$$

where θ_t is an unobserved linear Gaussian process and σ^2 is a fixed unknown parameter.

4.1 Basic SV model

In the basic SV model (2) the log-volatility process θ_t is assumed to be the Gaussian AR process as given by $\theta_t = h_t$, where

$$h_{t+1} = \phi_1 h_t + \ldots + \phi_p h_{t+1-p} + \xi_t, \qquad \xi_t \sim \mathrm{NID}(0, \sigma_\xi^2), \qquad t = 1, \ldots, n, \tag{29}$$

and with $p = 1$. The AR(1) process for h_t is given by (29) with $p = 1$ and is assumed to be independent of the Gaussian innovation sequence for ε_t. The mean is fixed, that is, $\mu_t = \mu$. Alternative specifications for μ_t and θ_t are discussed below.

The quasi-maximum likelihood analysis of the basic SV model is discussed in Section 2. It is made clear that such an analysis can be fully based on standard state-space methods applied to a linear Gaussian model. The transformation $y_t^* = \log(y_t - \bar{y})^2$ as in (5), for $t = 1, \ldots, n$, leads to the linear model (6) for $x_t = y_t^*$, where the disturbance $u_t + \kappa_1 = \log \varepsilon_t^2$ is $\log \chi^2$-distributed. The quasi-likelihood approach then replaces the $\log \chi^2$ by a Gaussian density for u_t with mean and variance equal to those of the $\log \chi^2$ density. However, Sandmann and Koopman (1998) considered model (6) as a linear model with the non-Gaussian $\log \chi^2$ density given by

$$\log p(y_t^*|\theta_t) = -\frac{1}{2}\log 2\pi + \frac{1}{2}\left(z_t - \exp z_t\right), \qquad z_t = y_t^* - \theta_t, \qquad t = 1, \ldots, n,$$

where $\theta_t = h_t$. We can adopt the Monte Carlo likelihood analysis of Section 2. The methods presented are fairly easy to implement in this case. The posterior mode calculations require expressions for

$$\dot{p}(y_t^*|\theta_t) = \frac{\partial}{\partial \theta_t}\log p(y_t^*|\theta_t), \quad \ddot{p}(y_t^*|\theta_t) = \frac{\partial^2}{\partial \theta_t \partial \theta_t'}\log p(y_t^*|\theta_t), \quad t = 1, \ldots, n.$$

In the case of the $\log \chi^2$ density $p(y_t^*|\theta_t)$, we have

$$\dot{p}(y_t^*|\theta_t) = \frac{1}{2}\left(\exp z_t - 1\right), \qquad \ddot{p}(y_t^*|\theta_t) = -\frac{1}{2}\exp z_t, \qquad t = 1, \ldots, n.$$

It follows that we can obtain the posterior mode via the Newton–Raphson updating steps (24) with the tth element of x given by $x_t = g_t + 1 - \exp(-z_t)$ and the tth diagonal element of A given by $A_t = 2\exp(-z_t)$. Once the posterior mode $\widetilde{\theta}$ has been obtained, simulation from the multivariate normal density $\mathrm{N}(\widetilde{\theta}, V)$ can take place via the simulation smoother and the Monte Carlo likelihood estimator (27) can be computed as a result. Further, it can be numerically maximized with respect to parameter vector ψ.

However, it is preferred to apply the methods of Section 3 directly on the SV model without transforming the log returns; see Shephard and Pitt (1997). In this case we need to treat the nonlinear observation model with a Gaussian density for ε_t. The conditional density $p(y_t|\theta_t)$ of the basic SV model is given by (3) with $h_t = \theta_t$. The posterior mode can be obtained as described in the previous paragraph but the derivatives of the model density are different and are given by

$$\dot{p}(y_t|\theta_t) = \frac{1}{2}\left\{\exp(-\theta_t)(y_t - \mu)^2 - 1\right\}, \qquad \ddot{p}(y_t|\theta_t) = -\frac{1}{2}\exp(-\theta_t)(y_t - \mu)^2,$$

for $t = 1, \ldots, n$. The methods for a Monte Carlo likelihood analysis can be implemented in a way similar to what has just been described.

4.2 Multiple volatility factors

In empirical work it has been observed that volatility often exhibits long-range dependence; see Andersen et al. (2003). Ideally, log volatility θ_t is modeled by a fractionally integrated process; see Granger and Joyeau (1980). Inference for the SV model (28) with a long-memory process for θ_t is often based on the spectral likelihood function; see Breidt et al. (1998) and Ray and Tsay (2000). Exact maximum likelihood methods were recently considered by Brockwell (2007). In our framework, we can approximate the long-range

dependence in the log volatility θ_t by considering it as a sum of independent autoregressive factors, that is,

$$\theta_t = \sum_{i=1}^{q} h_{it},$$

where each h_{it} represents the autoregressive process (29). The most commonly used specification is the two-factor model ($q = 2$), where one factor models the long-run dependence and the other the short-run dependence. The details of a Monte Carlo likelihood analysis are not different from those for a basic SV model since $p(y_t|\theta_t)$ remain unaltered.

4.3 Regression and fixed effects

It is often desirable to include regression effects in the specification of the volatility. Tsiakas (2006) introduced dummy effects to account for a seasonal pattern in the volatility of his periodic SV model. Koopman et al. (2005) considered a regression variable that contains information on the unobserved log-volatility process. Such regression effects can be incorporated into the SV model by extending the log-volatility signal specification by

$$\theta_t = W_t^{\theta}\beta + h_t,$$

where h_t is the autoregressive process (29), W_t^{θ} is a $1 \times k^{\theta}$ vector of covariates and β is a $k^{\theta} \times 1$ vector of regression coefficients.

The estimation of regression effects in the volatility process can be carried out in two ways. First, the coefficients in β can be treated as unknown parameters and incorporated in the parameter vector ψ. The Monte Carlo likelihood function (27) is maximized with respect to ψ that includes β. The methods of Section 3 can be applied straightforwardly. Second, the state vector α_t in the signal model (9) can be augmented so that it includes the coefficients in β. In this case (9) becomes

$$\alpha_t = \begin{pmatrix} h_t \\ \beta \end{pmatrix}, \quad \theta_t = (1, W_t^{\theta})\alpha_t, \quad \alpha_{t+1} = \begin{bmatrix} \phi & 0 \\ 0 & I \end{bmatrix} \alpha_t + \begin{pmatrix} \xi_t \\ 0 \end{pmatrix}, \quad Q_t = \begin{bmatrix} \sigma_{\xi}^2 & 0 \\ 0 & 0 \end{bmatrix},$$

for $t = 1, \ldots, n$, where h_t is taken here as an AR process of order $p = 1$. The initial variance matrix of the state vector P is block-diagonal with the two blocks given by $\sigma_{\xi}^2 \,/\, (1-\phi^2)$ and qI, where $q \to \infty$ and I is the $k^{\theta} \times k^{\theta}$ identity matrix. The diffuse prior for β reflects that β is fixed and unknown. State-space methods can be adjusted to handle diffuse prior conditions exactly; see Chapter 5 in Durbin and Koopman (2001). The Monte Carlo likelihood analysis of Section 3 also applies in this case. Depending on which way β is

estimated, the estimates will not be the same since the conditional likelihood approach produces a Monte Carlo maximum likelihood estimate, while the marginal likelihood approach produces a Monte Carlo conditional mean of β given the observations y; however, the differences are likely to be small. A similar discussion applies to linear Gaussian state-space models; see de Jong (1991).

Finally, regressors for the expected return can be included by the specification

$$\mu_t = \mu + W_t^{\mu}\delta,$$

where μ is an unknown fixed constant, W_t^{μ} is a $1 \times k^{\mu}$ vector of covariates and δ is a $k^{\mu} \times 1$ vector of regression coefficients. In this case we can regard model (28) as a regression model with SV errors. We include μ and δ in ψ and estimate ψ by Monte Carlo maximum likelihood using the methods of Section 3.

4.4 Heavy-tailed innovations

The excess kurtosis found in financial time series is often larger than can be explained by the basic SV model. This is caused by the fact that the excess kurtosis in the SV model is generated solely by randomness in the volatility. The model can be generalized by assuming that the innovations ε_t have a scaled t distribution. In this way, the dynamic properties of log volatility and and the thickness of tails are modeled separately. Examples of this approach can be found in Fridman and Harris (1998), Liesenfeld and Jung (2000) and Lee and Koopman (2004). We consider the SV model (28) with a scaled t distribution for ε_t. Its observation density is given by

$$\log p(y_t|\theta_t) = \log \frac{\Gamma(\frac{\nu}{2} + \frac{1}{2})}{\Gamma(\frac{\nu}{2})} - \frac{1}{2}\left\{\log \sigma^2(\nu-2) + \theta_t + (\nu+2)\log\left(1 + \frac{y_t^2}{(\nu-2)\sigma_t^2}\right)\right\},$$

for $t = 1, \ldots, n$, while the first and the second derivative with respect to θ_t are given by

$$\dot{p}(y_t|\theta_t) = \frac{1}{2} + \frac{1}{2}\frac{(\nu+1)y_t^2}{(\nu-2)\sigma_t^2 + y_t^2}, \qquad \ddot{p}(y_t|\theta_t) = -\frac{1}{2}\frac{\sigma_t^2(\nu-2)(\nu+1)y_t^2}{\{(\nu-2)\sigma_t^2 + y_t^2\}^2},$$

for $t = 1, \ldots, n$. The further details of a Monte Carlo likelihood analysis for this model are not different from those provided for a basic SV model in Section 3.

4.5 Additive noise

The basic SV model assumes that there is only one source of error. The SV model with additive noise assumes that there is an additional Gaussian noise term with constant variance, more specifically

$$\mu_t = \zeta_t, \qquad \zeta_t \sim \text{NID}(0, \sigma_\zeta^2).$$

This model was used in the context of high-frequency returns in Jungbacker and Koopman (2005). In the special case $\sigma_\zeta^2 = 0$ this model reduces to the basic SV model. The observation density is given by

$$\log p(y_t|\theta_t) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\left(\sigma_\zeta^2 + \sigma_t^2\right) - \frac{1}{2}\frac{y_t^2}{(\sigma_\zeta^2 + \sigma_t^2)},$$

while the first and the second derivative with respect to θ_t are given by

$$\dot{p}(y_t|\theta_t) = \frac{\sigma_t^2}{2(\sigma_t^2 + \sigma_\zeta^2)}\left\{\frac{y_t^2}{(\sigma_t^2 + \sigma_\zeta^2)} - 1\right\}$$

and

$$\ddot{p}(y_t|\theta_t) = \frac{\sigma_t^2}{2(\sigma_t^2 + \sigma_\zeta^2)}\left\{\frac{\sigma_t^2 + y_t^2}{(\sigma_t^2 + \sigma_\zeta^2)} - \frac{\sigma_t^2 y_t^2}{(\sigma_t^2 + \sigma_\zeta^2)^2} - 1\right\},$$

for $t = 1, \ldots, n$, respectively. Since $\ddot{p}(y_t|\theta_t)$ is not necessarily negative for all $t = 1, \ldots, n$, the observation density is not necessarily log-concave. We therefore need to rely on the arguments of Jungbacker and Koopman (2007) to carry out a Monte Carlo likelihood analysis. However, the methods and techniques have not changed intrinsically and the descriptions in Section 3 still apply.

4.6 Leverage effects

The leverage effect occurs if a negative return increases the volatility more than a positive return of the same magnitude decreases it; see the seminal paper of Black (1976) where this phenomenon was described originally. The leverage effect is incorporated in the SV model by allowing correlation between the innovations of the state and the observation equation; see Yu (2005) for a detailed discussion. The SV model with leverage and based on an AR(1) process for log volatility is given by

$$y_t = \sigma \exp(\frac{1}{2}h_t)\varepsilon_t, \qquad h_{t+1} = \phi h_t + \xi_t, \qquad \begin{pmatrix}\varepsilon_t \\ \xi_t\end{pmatrix} \sim \text{NID}\left(0, \begin{bmatrix}1 & \sigma_\xi\rho \\ \sigma_\xi\rho & \sigma_\xi^2\end{bmatrix}\right),$$

for $t = 1, \ldots, n$. The correlation coefficient ρ is typically negative, implying that negative shocks in the return are accompanied by positive shocks in the volatility and vice versa.

The general formulation of the SV model with leverage requires both h_t and ξ_t with $t = 1, \ldots, n$ in θ since h_t appears directly in the observation equation and ξ_t is required to measure the correlation with ε_t. The variance matrix Ψ of the signal θ in (12) is therefore singular and the methods of Section 3 clearly break down. We treat this problem by following Jungbacker and Koopman (2007) and reformulate the model by

$$y_t = \sigma \exp(\frac{1}{2}h_t^*) \left\{\varepsilon_t^* + \text{sign}(\rho)\xi_{2t}\right\}, \qquad \varepsilon_t^* \sim \text{NID}(0, 1 - |\rho|),$$

where

$$h_{t+1}^* = \phi h_t^* + \sigma_\xi \left(\xi_{1,t} + \xi_{2t}\right), \quad \xi_{1t} \sim \text{NID}(0, 1 - |\rho|), \quad \xi_{2t} \sim \text{NID}(0, |\rho|),$$

for $t = 1, \ldots, n$, with $h_1^* \sim N\{0, \sigma_\xi^2(1 - \phi^2)^{-1}\}$. The disturbances ε_t^*, ξ_{1t} and ξ_{2t} are mutually and serially independent disturbances for $t = 1, \ldots, n$. The signal vector for this model formulation contains h_t^* and ξ_{2t} with $t = 1, \ldots, n$ and the corresponding variance matrix Ψ is nonsingular as required. In terms of the general formulation (9), we have $\alpha_t = (h_t^* \ , \ \sigma_\xi \xi_{2,t})'$, $\xi_t = \sigma_\xi(\xi_{1,t} \ , \ \xi_{2,t+1})'$ and

$$\theta_t = \alpha_t, \quad \alpha_{t+1} = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} \alpha_t + \xi_t, \quad \begin{aligned} \xi_t &\sim \text{NID}\left\{0, \sigma_\xi^2 \text{diag}(1 - |\rho|, |\rho|)\right\}, \\ \alpha_1 &\sim \text{NID}\left\{0, \sigma_\xi^2 \text{diag}([1 - \phi^2]^{-1}, |\rho|)\right\}, \end{aligned}$$

for $t = 1, \ldots, n$. The observations $y_1, \ldots, y_n$ have the conditional density $\log p(y|\theta) = \sum_{t=1}^n \log p(y_t|\theta_t)$, where

$$\log p(y_t|\theta_t) = c - \frac{1}{2}h_t^* - \frac{1}{2}\sigma^{-2} \exp(-h_t^*)(1-|\rho|)^{-1}\{y_t - \sigma \exp(\frac{1}{2}h_t^*)\text{sign}(\rho)\xi_{2,t}\}^2,$$

for $t = 1, \ldots, n$, where c is some constant. Expressions for the 2×1 vector $\dot{p}(y_t|\theta_t)$ and the 2×2 matrix $\ddot{p}(y_t|\theta_t)$, as defined in (22), can be obtained straightforwardly. It turns out that density $p(y|\theta)$ is not necessarily log-concave. The Monte Carlo estimator (27) of the likelihood function can be evaluated by using the methods of Section 3 and by adopting the arguments of Jungbacker and Koopman (2007), who also presented an empirical illustration.

4.7 Stochastic volatility in mean

As investors require a larger expected return if the risk is large, it seems reasonable to expect a positive relationship between volatility and returns. Empirical evidence however points to a negative influence of volatility on returns; see French et al. (1987). This effect can be explained by assuming a positive relationship between expected return and *ex ante* volatility. Koopman and Hol-Uspensky (2002) proposed capturing this so-called volatility feedback effect by including volatility as a regression effect in the mean:

$$\mu_t = a + by_{t-1} + d\sigma \exp\left(\frac{1}{2}\theta_t\right),$$

where a, b, d and σ^2 are parameters. We assume that both mean and variance processes are stationary. The volatility feedback effect coefficient d is typically negative, if not zero.

The observation density is given by

$$\log p(y_t|\theta_t) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2}h_t - \frac{\left(y_t - a - by_{t-1} - d\sigma_t^2\right)^2}{2\sigma_t^2},$$

while the first and the second derivative with respect to θ_t are given by

$$\dot{p}(y_t|\theta_t) = -\frac{1}{2} + \left(y_t - a - by_{t-1} - d\sigma_t^2\right)d + \frac{\left(y_t - a - by_{t-1} - d\sigma_t^2\right)^2}{2\sigma_t^2}$$

and

$$\ddot{p}(y_t|\theta_t) = -\frac{1}{2} - d^2\sigma_t^2 - \dot{p}(y_t|\theta_t),$$

for $t = 1, \ldots, n$, respectively. It can be shown that density $p(y|\theta)$ is log-concave, so $\ddot{p}(y_t|\theta_t) < 0$ for all $t = 1, \ldots, n$. The further details of a Monte Carlo likelihood analysis for this model are therefore not different from those provided for a basic SV model in Section 3.

5 Empirical Illustrations

Three financial time series are analyzed to illustrate the methods presented in this chapter. The daily returns of the Standard & Poor's 500 (S&P500) stock index (January 3, 1991 to October 20, 2006: 3,985 observations, weekends and holidays excluded), the daily changes in the US dollar–pound sterling exchange rates and the daily changes in the US dollar–Japanese yen exchange rates (both for January 3, 1990 to October 20, 2006: 4,383 observations,

weekends excluded but with missing values for holidays). These time series are obtained from Datastream and are presented graphically in Fig. 1.

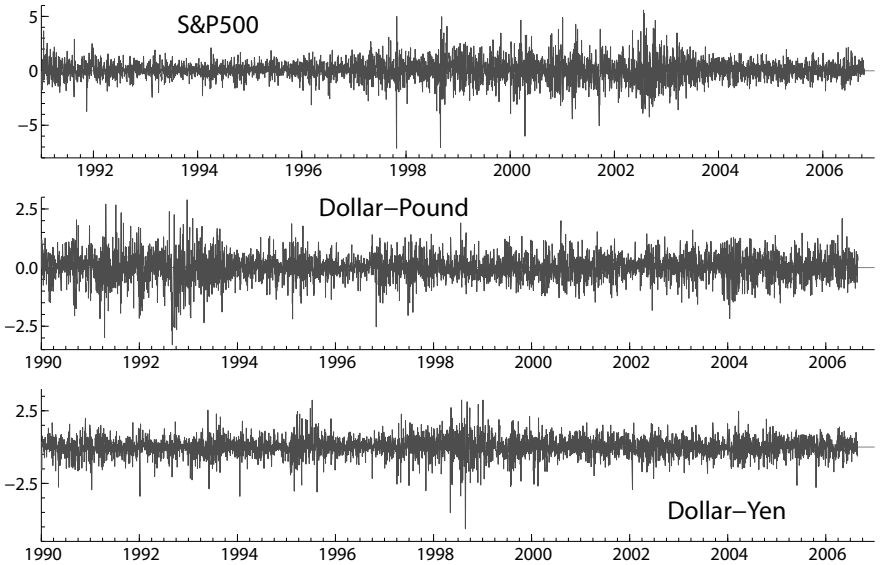

Fig. 1 Daily returns of S&P500 and daily changes of dollar–pound and dollar–yen exchange rate series

5.1 Standard & Poor's 500 stock index: volatility estimation

The volatility in the S&P500 returns is subject to some breaks: low volatility before 1997, high volatility between 1997 and 2004 with peaks in the aftermath of September 11, 2001 and moderate volatility from the end of 2003. Further, the clustering of periods with low and high volatility is clearly visible in the series. An exploratory data analysis is carried out firstly and the STAMP program of Koopman et al. (2007) is used for the analysis, estimation and signal extraction on the basis of the linearized SV model. The estimation results and the estimated volatility are indicative of the salient features in the volatility of S&P500 returns.

The Monte Carlo likelihood methods of Section 3 are used for the estimation of parameters with $M = 100$. Various SV models are considered for this purpose. First, the parameters in the SV models with log volatility modeled by a single AR(1) term (basic SV) and by two AR(1) terms (multiple volatility factors) are estimated by Monte Carlo maximum likelihood and the

results are presented in the second and third columns of Table 1. The standard errors of the estimates are based on the information matrix, which is evaluated numerically. The estimated coefficients are reported together with their standard errors in parentheses and it appears that for both models the estimates are significantly different from zero. However, the estimated autoregressive coefficients are close to unity and it is known that standard errors for parameter estimates close to boundary values are not reliable; see also the extensive literature on unit root testing. However, the likelihood ratio test value for the inclusion of a second AR(1) term in the SV model with multiple volatility factors is 18.8 and has a p value of 0.0001. The estimated volatilities are presented in Fig. 2. The main patterns reflect the salient features of the dynamics in volatility. The volatility increase in 1997, the peak in the early years of the twenty-first century and the slowdown in 2003 can be observed clearly from the estimated volatility patterns. The distinct difference between the estimated volatility patterns of the one- and two-factor volatilities is that the signal is noisier for the multiple-factor SV model; however, the main patterns are similar. Given the likelihood improvement, the noisier volatility estimate for the two-factor model appears to provide a better local estimate of the underlying volatility.

The basic SV model with a t density for ε_t is also considered for the S&P500 series. The parameter estimates can be found in the fourth column of Table 1. It is expected that the large shocks in the return series will be captured more effectively in the tails of the t density and therefore the estimated volatility is more persistent and smoother. The estimation results confirm this. The estimated value for ϕ is 0.995, which is slightly larger than the value of 0.991 for the basic SV model. Furthermore, the estimate of σ_ξ^2 (the volatility of volatility) has a smaller value for the SV model with t density than for the basic SV model and hence the estimated volatility is smoother. This is confirmed by the bottom graph in Fig. 2. The estimated underlying volatility pattern for the S&P500 return series is smoothest for the SV model with t density.

5.2 Standard & Poor's 500 stock index: regression effects

SV models can incorporate regression effects in different ways as shown in Section 4. The SV in the mean model is considered to investigate the feedback between returns and volatility. For this purpose we consider daily excess returns for the S&P500 stock index series where excess return is defined as the return minus the risk-free return. The estimation is carried out using the Monte Carlo likelihood methods for which some details are given in Section 4. The estimation results are presented in the final column of Table 1. Apart from the basic SV parameters, the parameters of interest are a, b and d,

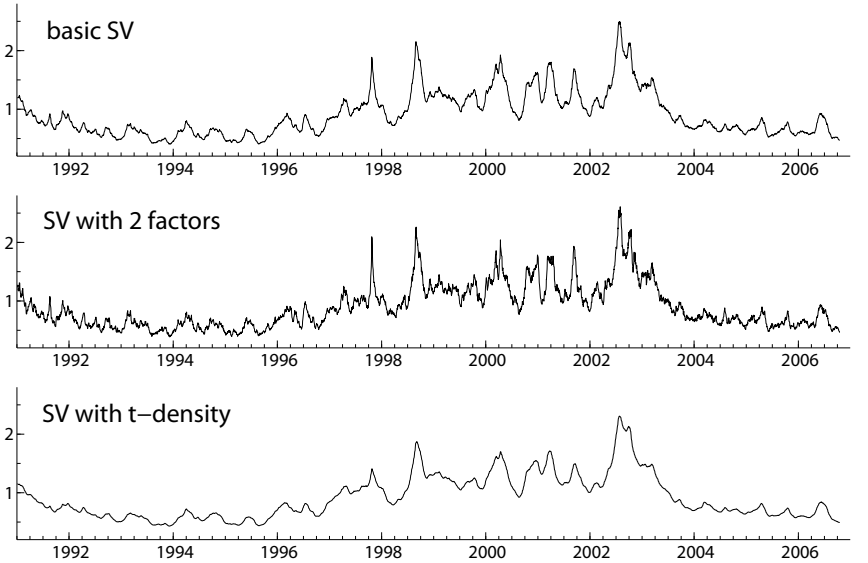

Fig. 2 Estimated volatilities σ_t for S&P500 returns and for three stochastic volatility (SV) models: basic, with two factors and with t density

Table 1 Monte Carlo maximum likelihood estimates of parameters (with standard errors) for stochastic volatility (SV) models applied to S&P 500 daily returns (SV-t: SV model with t density; for the SV in mean model applied to excess returns, the deviations are from risk-free returns)

Coefficients	Basic SV	SV–2 factors	SV-t	SV in mean
ϕ	0.991 (0.003)	0.998 (0.001)	0.995 (0.002)	0.976 (0.006)
σ_ξ^2	0.0127 (0.003)	0.0014 (0.0008)	0.006 (0.00186)	0.028 (0.006)
ϕ	—	0.93 (0.029)	—	—
σ_ξ^2	—	0.028 (0.01)	—	—
σ^2	0.681 (0.129)	0.643 (0.227)	0.695 (0.17401)	0.672 (0.072)
ν	—	—	10.651 (2.1058)	—
a	—	—	—	0.072 (0.018)
b	—	—	—	0.051 (0.016)
d	—	—	—	−0.038 (0.025)

which are estimated significantly except parameter d; however, the estimate of d is not too far away from its significance level and it is a negative value. This confirms the empirical findings of French et al. (1987) and Koopman and Hol-Uspensky (2002), among others.

An alternative volatility indicator is the daily price range defined as $Z_t^\beta = \log(\max(\log Pr_t) - \min(\log Pr_t))$ where Pr_t is the vector of realized stock prices within a certain day t. This daily S&P500 stock index price range is also indicative of the amount of volatility and can be used to detect volatility in daily returns. We therefore consider the SV model with $\theta = h_t + Z_t^\beta \beta$, the estimation results for which are presented in Table 2. Although similar results are obtained as for the SV models without the inclusion of the covariate Z_t^β in the signal θ_t, the estimates for β are rather mixed. The SV model with one volatility factor produces a highly significant and positive estimate for β, the model with two volatility factors produces a highly significant but negative estimate for β, while the SV model with a t density produces an insignificant estimate for β. Further research and empirical evidence is required to assess the increased performance of incorporating range-based volatility measures in SV models.

Table 2 Monte Carlo maximum likelihood estimates of parameters (with standard errors) from *SVX* models for S&P 500 returns

Coefficients	Basic SV	SVX	SVX–2 factors	SVX-t
ϕ	0.991 (0.003)	0.991 (0.0029)	0.998 (0.0009)	0.996 (0.0019)
σ_ξ^2	0.0127 (0.003)	0.0124 (0.003)	0.0014 (0.0008)	0.006 (0.0017)
ϕ	—	—	0.929 (0.029)	—
σ_ξ^2	—	—	0.031 (0.011)	—
σ^2	0.681 (0.129)	0.71 (0.234)	0.564 (0.267)	0.911 (0.256)
ν	—	—	—	10.541 (0.351)
β	—	0.0156 (0.003)	−0.033 (0.005)	0.002 (0.002

5.3 Daily changes in exchange rates: dollar–pound and dollar–yen

For the analysis of volatility in exchange rate series, the interest focuses solely on the signal extraction of volatility. For this purpose we have considered the exchange rates for pound sterling and Japanese yen, both against the US dollar. Three different SV model specifications are applied to the two daily change series. The first model is the basic SV model, the second is the SV

model with additive noise and the third is the SV model with a t density for the model equation. In all cases the estimation methods of Section 3 are used. In the case of the SV with noise model, the observation density $p(y|\theta)$ is not log-concave and the recent modifications need to be applied. The implementation in all three cases has been successful and the estimation results are presented in Table 3. The persistency of the volatility clearly increases for an SV model with a t density, while the additive noise does not seem to affect the dynamic properties of volatility. Although the additive noise is highly significant for the dollar–yen series, it is not significant for the dollar–pound series. This empirical finding may be explained by factors such as trading volumes and information flows.

Table 3 Monte Carlo maximum likelihood estimates of parameters (with standard errors) from SV models for daily changes in exchange rates (*SV-t*: SV model with t density; *SVN*: SV model with additive noise

	Dollar–pound			Dollar–yen		
Coefficients	Basic SV	SVN	SV-t	Basic SV	SVN	SV-t
ϕ	0.977 (0.00606)	0.977 (0.00609)	0.986 (0.00435)	0.934 (0.02)	0.933 (0.02)	0.987 (0.0041)
σ_ξ^2	0.0179 (0.00477)	0.0200 (0.0075)	0.00976 (0.00294)	0.0521 (0.018)	0.115 (0.037)	0.0064 (0.002)
σ^2	0.265 (0.0245)	0.248 (0.0487)	0.274 (0.0306)	0.371 (0.0222)	0.224 (0.031)	0.404 (0.041)
σ_ζ^2	—	0.0144 (0.0348)	—	—	0.115 (0.021)	—
ν	—	—	10.459 (1.843)	—	—	6.561 (0.681)

The estimated volatilities obtained from the three models and the two exchange series are presented in Fig. 3 for the dollar–pound series and in Fig. 4 for the dollar–yen series. The salient features of the volatility for both series are clearly captured and it is interesting that the volatility patterns for both series are distinct from each other. In the early years of the 1990s, the dollar–pound series is subject to higher volatility. After 1994, the volatility is moderate for 10 years but increases somewhat from 2004. Throughout the 1990s, the volatility is relatively high for the dollar–yen series and the volatility of volatility is also high. A clearly high period of volatility occurred during the Asian financial crises in the late 1990s. However, in the early years of the new millennium, the volatility stabilizes and seems to behave more in par with the dollar–pound series. This may be indicative of the convergence of international financial markets that is discussed in the economics and finance literature.

The differences in the estimated volatilities for the three models are similar for both exchange series. The volatility patterns for the SV model with noise also turn out to be noisier than those obtained for the basic SV model. The smoothed patterns of volatility are obtained from the SV model with a t density. Given the estimation results presented in Table 3, the estimated

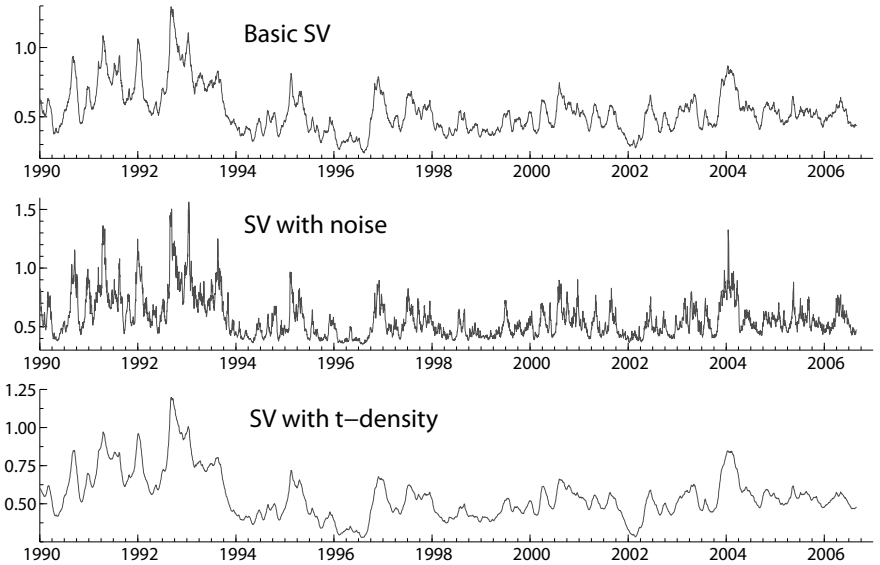

Fig. 3 Volatility estimates for daily changes in dollar–pound exchange rates

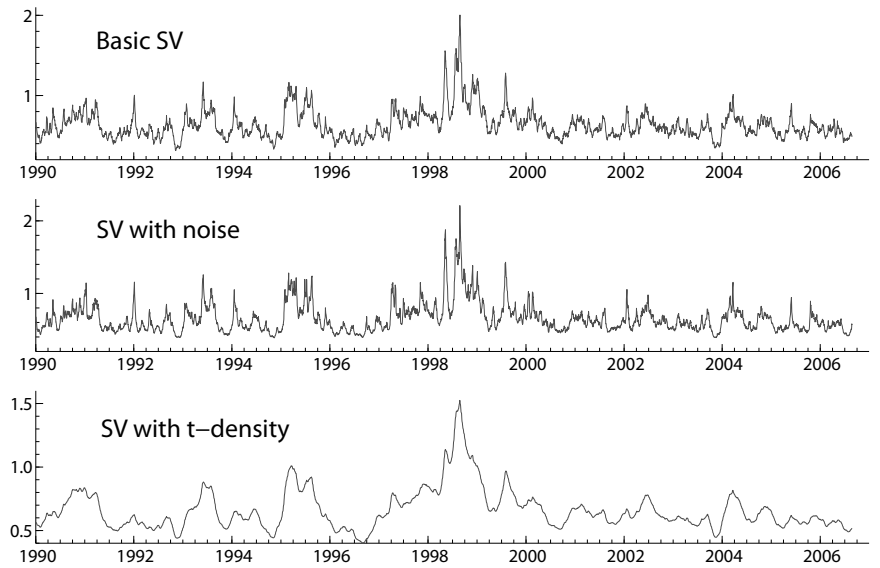

Fig. 4 Volatility estimates for daily changes in dollar–yen exchange rates

volatility patterns confirm that the different SV models capture different features of volatility in the series.

6 Conclusions

This chapter has reviewed parameter estimation methods for the basic SV model. We have restricted ourselves to estimation methods based on the linearization of the SV model and methods based on Monte Carlo simulations. In the former case, fast and linear estimation methods for the standard linear Gaussian state-space model can be adopted, such as the Kalman filter and the associated smoothing algorithm. Classical maximum likelihood methods are used for parameter estimation and for which standard software tools are available. In the latter case, Monte Carlo simulations are used for the evaluation of the likelihood function that is expressed as an integral. A convenient analytical expression is not available and therefore one needs to rely on numerical methods. Importance sampling methods have been suggested as a feasible way to evaluate the likelihood function so that it can be maximized numerically with respect to a set of parameters. The details of this approach were discussed in this chapter and particularly in Section 3. This chapter further reviewed some interesting extensions of the SV model, including models with explanatory variables, with additive noise, with leverage and with a t density for the observation model. It was shown that parameters in these more general SV models can also be estimated by the Monte Carlo methods discussed in this chapter. The empirical results illustrate that the general SV models and the associated methods can successfully capture interesting aspects of volatility.

Acknowledgement We gratefully thank the editors for their insightful comments on an earlier draft of this chapter. All remaining errors are ours.

Appendix: State-Space Methods

In this chapter we consider the linear Gaussian state space model with

- State equation:

$$\alpha_{t+1} = d_t + T_t\alpha_t + \eta_t, \qquad \eta_t \sim \mathrm{NID}(0, Q_t).$$

- Initial condition: $\alpha_1 \sim \mathrm{N}(a, P)$.
- Signal equation:

$$\theta_t = c_t + Z_t\alpha_t.$$

- Observation equation:

$$x_t = \theta_t + u_t, \qquad u_t \sim \text{NID}(0, H_t), \qquad t = 1, \ldots, n.$$

The system vectors c_t and d_t and system matrices Z_t, H_t, T_t and Q_t are fixed and known functions of parameter vector ψ. The observation x_t and signal θ_t are assumed to be scalar variables, while the state vector α_t together with the disturbance vector η_t have dimensions $q \times 1$. All state-space quantities have appropriate dimensions.

Kalman filter

The Kalman filter equations are given by

$$\begin{aligned} v_t &= x_t - c_t - Z_t a_t, & F_t &= H_t + Z_t P_t Z_t', \\ & & K_t &= T_t P_t Z_t' F_t^{-1}, & t &= 1, \ldots, n, \\ a_{t+1} &= d_t + T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t T_t' - K_t F_t K_t' + Q_t, \end{aligned} \tag{30}$$

for $t = 1, \ldots, n$. The Kalman filter is initialized by $a_1 = a$ and $P_1 = P$, where a and P are the mean and variance of the initial state vector α_1, respectively. The quantities of the Kalman filter are usually stored so they can be used for other purposes such as smoothing. The Kalman filter is a forwards recursion. The one-step-ahead prediction errors v_t and their variances F_t are used for the evaluation of the Gaussian likelihood function and as residuals for diagnostic checking.

Smoothing algorithm

Once the Kalman filter has been carried out and the quantities K_t, F_t and v_t have been stored for $t = 1, \ldots, n$, the smoothing recursions enable the estimation of the smoothed estimate of θ_t and its variance matrix given by $\widehat{\theta}_t = E(\theta_t | y)$ and $V_t = E(\theta_t - \widehat{\theta}_t)(\theta_t - \widehat{\theta}_t)'$. The recursions operate backwards and are given by

$$\begin{aligned} \widehat{\theta}_t &= x_t - H_t u_t, & V_t &= H_t D_t H_t, \\ u_t &= F_t^{-1} v_t - K_t' r_t & D_t &= F_t^{-1} - K_t' N_t K_t, \\ r_{t-1} &= Z_t' u_t + T_t' r_t, & N_{t-1} &= Z_t' D_t Z_t + T_t' N_t T_t - Z_t' K_t' N_t T_t - T_t' N_t K_t Z_t, \end{aligned} \tag{31}$$

for $t = n,\ n-1, \ldots, 1$ and with the initializations $r_n = 0$ and $N_n = 0$.

Simulation smoothing algorithm

Once the Kalman filter has been carried out and the quantities K_t, F_t and v_t have been stored for $t = 1, \ldots, n$, the simulation smoothing recursions enable the generation of draws from the density $f(\theta; y)$. The recursions operate backwards as in smoothing and are given by

$$\begin{aligned} C_t = B_t B_t' = H_t^{-1} - F_t^{-1} - K_t' N_t K_t, \quad & R_t = C_t^{-1}(H_t^{-1} Z_t - K_t' N_t T_t), \\ o_t \sim \mathrm{N}(0, I), \qquad w_t \;=\; B_t o_t, \qquad & u_t = H_t(w_t + F_t^{-1} v_t - K_t' r_t), \\ r_{t-1} = Z_t' H_t^{-1} u_t - R_t' w_t + T_t' r_t, \qquad & N_{t-1} = R_t' C_t R_t - Z_t' H_t^{-1} Z_t + T_t' N_t T_t, \end{aligned} \tag{32}$$

for $t = n,\ n-1, \ldots, 1$ and with the initializations $r_n = 0$ and $N_n = 0$. The simulation θ^i from $p(\theta|y)$ is computed by $\widehat{\theta} + (u_1', \ldots, u_n')'$, where $\widehat{\theta} = E(\theta|y)$ and is obtained from the smoothing algorithm. Given the simulation θ^i, the importance density function $f(\theta^i; y)$ in (27) can be evaluated by

$$f(\theta^i; y) = \exp\left(-\frac{mn}{2} \log 2\pi - \sum_{t=1}^{n} \log |H_t| - \sum_{t=1}^{n} \log |B_t| - \frac{1}{2} \sum_{t=1}^{n} o_t^i \, {}' o_t^i \right). \tag{33}$$

References

Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. (2003): Modelling and forecasting realized volatility. *Econometrica* **71**, 529–626.

Andersen, T. G. and Sorensen, B. (1996): GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Business and Economic Statistics* **14**, 328–352.

Black, F. (1976): Studies of stock price volatility changes. *Proceedings of the Business and Economic Statistics Section*, 177–181.

Bollerslev, T. (1986): Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* **51**, 307–327.

Breidt, F. J. and Carriquiry, A. L. (1996): Improved quasi-maximum likelihood estimation for stochastic volatility models. *In: Jack, W. O. J., Lee, C. and Zellner, A. (Eds.): Modelling and Prediction: Honoring Seymour Geisser*. Springer, New York, 228–247.

Breidt, F. J., Crato, N. and de Lima, P. (1998): On the detection and estimation of long memory in stochastic volatility. *Journal of Econometrics* **83**, 325–348.

Brockwell, A. E. (2007): Likelihood-based analysis of a class of generalized long-memory time series models. *Journal of Time Series Analysis* **28**, 386–407.

Davis, R. A. and Rodriguez-Yam, G. (2005): Estimation for state-space models based on a likelihood approximation. *Statistica Sinica* **15**, 381–406.

de Jong, P. (1991): The diffuse Kalman filter. *Annals of Statistics* **19**, 1073–83.

de Jong, P. and Shephard, N. (1995): The simulation smoother for time series models. *Biometrika* **82**, 339–50.

Durbin, J. and Koopman, S. J. (1997): Monte Carlo maximum likelihood estimation of non-Gaussian state space model. *Biometrika* **84**, 669–84.

Durbin, J. and Koopman, S. J. (2001): *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Durbin, J. and Koopman, S. J. (2002): A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**, 603–16.

Engle, R. F. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica* **50**, 987–1007.

French, K. R., Schwert, G. W. and Stambaugh, R. F. (1987): Expected stock returns and volatility. *J. Financial Economics* **19**, 3–29. Reprinted as pp. 61–86 in Engle, R.F.(1995): *ARCH: Selected Readings*. Oxford University Press, Oxford.

Fridman, M. and Harris, L. (1998): A maximum likelihood approach for non-gaussian stochastic volatility models. *Journal Business and Economic Statistics* **16**, 284–291.

Fuller, W. A. (1996): *Introduction to Time Series* (2nd ed.) Wiley, New York.

Geweke, J. (1989): Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–39.

Granger, C. W. J. and Joyeau, R. (1980): An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–39.

Harvey, A. C., Koopman, S. J. and Penzer, J. (1998): Messy time series. *In: Fomby, T. B. and Hill, R. C. (Eds.): Advances in Econometrics* **13**, 103–143. JAI Press, New York.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994): Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–64.

Jacquier, E., Polson, N. and Rossi, P. (1994): Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business and Economic Statistics* **12**, 371–417.

Jungbacker, B. and Koopman, S. J. (2005): Model-based measurement of actual volatility in high-frequency data. *In: Fomby, T. B. and Terrell, D. (Eds.): Advances in Econometrics*. JAI Press, New York.

Jungbacker, B. and Koopman, S. J. (2007): Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika* **94**, 827–39.

Kim, S., Shephard, N. and Chib, S. (1998): Stochastic volatility: likelihood inference and comparison with arch models. *Review of Economic Studies* **65**, 361–393.

Koopman, S. J., Harvey, A. C., Doornik, J. A. and Shephard, N. (2007): *Stamp 8.0: Structural Time Series Analyser, Modeller and Predictor*. Timberlake Consultants, London.

Koopman, S. J. and Hol-Uspensky, E. (2002): The Stochastic Volatility in Mean model: Empirical evidence from international stock markets. *Journal of Applied Econometrics* **17**, 667–89.

Koopman, S. J., Shephard, N. and Doornik, J. A. (1999): Statistical algorithms for models in state space form using SsfPack 2.2. *Econometrics Journal* **2**, 113–66. http://www.ssfpack.com/.

Koopman, S. J., Jungbacker, B. and Hol, E. (2005): Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance* **12**, 445–75.

Koopman, S. J., Shephard, N. and Creal, D. (2007): Testing the assumptions behind importance sampling. *Discussion paper, VU University, Amsterdam*.

Lee, K. M. and Koopman, S. J. (2004): Estimating stochastic volatility models: a comparison of two importance samplers. *Studies in Nonlinear Dynamics and Econometrics* **8**, Art 5.

Liesenfeld, R. and Jung, R. (2000): Stochastic volatility models: conditional normality versus heavy-tailed distributions. *Journal of Applied Econometrics* **15**, 137–160.

Nocedal, J. and Wright, S. J. (1999): *Numerical Optimization* Springer, New York.

Ray, B. and Tsay, R. (2000): Long- range dependence in daily stock volatilities. *Journal of Business and Economic Statistics* **18**, 254–62.

Sandmann, G. and Koopman, S. J. (1998): Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics* **87**, 271–301.

Shephard, N. (2005): *Stochastic Volatility: Selected Readings*. Oxford University Press, Oxford.

Shephard, N. and Pitt, M. K. (1997): Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–67.

So, M. K. P. (2003): Posterior mode estimation for nonlinear and non-Gaussian state space models. *Statistica Sinica* **13**, 255–274.

Taylor, S. J. (1986): *Modelling Financial Time Series*. Wiley, Chichester.

Tsiakas, I. (2006): Periodic stochastic volatility and fat tails. *Journal of Financial Econometrics* **4**, 90–135.

Yu, J. (2005): On leverage in a stochastic volatility model. *Journal of Econometrics* **127**, 165–78.

Zivot, E. (2008): Practical aspects of GARCH-Modelling. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 112–155. Springer, New York.

Zivot, E., Wang, J. and Koopman, S. J. (2004): State space modeling in macroeconomics and finance using SsfPack in S+FinMetrics. *In: Harvey, A. C., Koopman, S. J. and Shephard, N. (Eds.): State space and unobserved components models*. Cambridge University Press, Cambridge.

Stochastic Volatility Models with Long Memory

Clifford M. Hurvich and Philippe Soulier

Abstract In this contribution, we consider models in discrete time that contain a latent process for volatility. The most well-known model of this type is the Long-Memory Stochastic Volatility (LMSV) model. We describe its main properties, discuss parametric and semiparametric estimation for these models, and give some generalizations and applications.

1 Introduction

In this contribution we consider models for long memory in volatility. There are a variety of ways to construct such models. Our primary focus here will be on models in discrete time that contain a latent process for volatility. The most well-known model of this type is the Long-Memory Stochastic Volatility (LMSV) model, proposed independently by Breidt, Crato and de Lima (1998) and Harvey (1998). It is a long-memory generalization of the Stochastic Volatility (SV) model of Taylor (1986). The LMSV model is appropriate for describing series of financial returns at equally-spaced intervals of time. The model implies that returns are a finite-variance martingale difference sequence, hence uncorrelated, while power transformations of the absolute returns have slowly decaying autocorrelations, in keeping with the empirical findings of Ding, Granger and Engle (1993). We will present the LMSV model, explain its basic properties, and give a survey of existing theoretical results. A variety of generalizations of the model have been considered, and

Clifford M. Hurvich
New York University, 44 W. 4'th Street, New York NY, 10012, USA, e-mail: churvich@stern.nyu.edu

Philippe Soulier
Mathematiques, Université Paris X–Nanterre, 200, Avenue de la République, F–92000 Nanterre, France, e-mail: philippe.soulier@u-paris10.fr

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_14,

some of these will be briefly discussed, but in order to enhance readability we will focus on a basic form of the model.

An important distinction between ARCH-type models and SV-type models is that the former are observation-driven, giving an expression for the one-step-ahead conditional variance in terms of observables and model parameters, while the latter are driven by a latent (unobserved) process which stands as a proxy for volatility but which does not represent the conditional variance. Thus, for the LMSV model it is necessary to use and develop appropriate techniques in order to carry out basic activities such as forecasting of squared returns, or aggregates of these (i.e., the realized volatility; see, e.g., Andersen, Bollerslev, Diebold and Labys (2001)), as well as estimation of parameters.

For simplicity, we will assume that the latent long-memory process is stationary and Gaussian, and is independent of the multiplying shock series (see Equation (1) below). We will consider parameter estimation, forecasting, smoothing, as well as semiparametric estimation and hypothesis testing for the long memory parameter. Besides presenting theoretical results, we will also discuss questions of computational efficiency.

There are several definitions of long memory, which are not equivalent in general (see Taqqu (2003)). For simplicity, we will say here that a weakly stationary process has long memory if its autocovariances $\{c_r\}$ satisfy

$$c_r \sim K_1 r^{2d-1}$$

($K_1 > 0$) as $r \to \infty$, or if its spectral density $f(\omega)$, $\omega \in [-\pi, \pi]$ satisfies

$$f(\omega) \sim K_2 |\omega|^{-2d}$$

($K_2 > 0$) as $\omega \to 0$, where $d \in (0, 1/2)$ is the memory parameter.

2 Basic Properties of the LMSV Model

The LMSV model for a stationary series of returns $\{r_t\}$ takes the form

$$r_t = \exp(Y_t/2) e_t \tag{1}$$

where $\{e_t\}$ is a series of i.i.d. shocks with zero mean, finite variance, and $\{Y_t\}$ is a zero-mean stationary Gaussian long-memory process, independent of $\{e_t\}$, with memory parameter $d \in (0, 1/2)$. Since $\{e_t\}$ is a martingale difference sequence, so is $\{r_t\}$. It follows that $\{r_t\}$ is a white noise sequence with zero mean. As is the case for most existing volatility models, the LMSV model is nonlinear, in that it cannot be represented as a linear combination of an i.i.d. series.

To study persistence properties of volatility, Ding, Granger and Engle (1993) used power transformations of absolute returns. Using the properties of the lognormal distribution, (see for example Harvey (1998), equation (12.9); cf. Robinson and Zaffaroni (1997) and (1998), Robinson (2001) it is possible to derive an explicit expression for the autocorrelations of $\{|r_t|^c\}$ for any positive c such that $\mathrm{E}[|e_t|^{2c}]$ is finite. The expression implies that the $\{|r_t|^c\}$ have long memory with the same memory parameter d, for all such c. It follows that if $\mathrm{E}[|e_t|^4]$ is finite and M is a fixed positive integer representing the degree of aggregation, the realized volatility $\{RV_k\}$ given by

$$RV_k = \sum_{t=(k-1)M+1}^{kM} r_t^2$$

has long memory with memory parameter d.

For estimation, it is convenient to work with the logarithms of the squared returns, $\{X_t\} = \{\log r_t^2\}$, which have the signal plus noise representation

$$X_t = \mu + Y_t + \eta_t, \tag{2}$$

where $\mu = \mathrm{E}[\log e_t^2]$ and $\{\eta_t\} = \{\log e_t^2 - \mathrm{E}[\log e_t^2]\}$ is an i.i.d. process independent of $\{Y_t\}$. Thus, the log squared returns $\{X_t\}$ are expressed as the sum of the long-memory process $\{Y_t\}$, the signal, and the i.i.d. process $\{\eta_t\}$, the noise, which is independent of the signal. It follows from (2) that the autocovariances of $\{X_t\}$ are equal to those of $\{Y_t\}$ for all nonzero lags. Therefore, the autocorrelations of $\{X_t\}$ are proportional to those of $\{Y_t\}$. Furthermore, the spectral density of $\{X_t\}$ is given by

$$f_X(\omega) = f_Y(\omega) + \sigma_\eta^2/(2\pi), \tag{3}$$

where $\sigma_\eta^2 = \mathrm{var}(\eta_t)$, assumed to be finite, and hence we have

$$f_X(\omega) \sim K_2|\omega|^{-2d}$$

($K_2 > 0$) as $\omega \to 0$. Thus, the log squared returns $\{X_t\}$ have long memory with memory parameter d.

3 Parametric Estimation

In both Harvey (1998) and Breidt, Crato and de Lima (1998) it is assumed that $\{Y_t\}$ is generated by a finite-parameter model. This model is taken to be the $ARFIMA(0, d, 0)$ model in Harvey (1998) and the $ARFIMA(p, d, q)$ model in Breidt, Crato and de Lima (1998). Given any finite-parameter long-memory specification for $\{Y_t\}$ in the LMSV model, we face the problem

of estimating the model parameters based on observations $r_1, \ldots, r_n$. Full maximum likelihood is currently infeasible from a computational point of view since it would involve an n-dimensional integral. Since long-memory models do not have a state-space representation, it is not possible to directly use a variety of techniques that have been successfully implemented for estimation of autoregressive stochastic volatility models (see, e.g., Harvey, Ruiz and Shephard (1994). We consider here two variants of Gaussian quasi maximum likelihood (QML), in the time and frequency domains. Both are based on the log squared returns, $\{X_t\}_{t=1}^n$, and both are presumably inefficient compared to the (infeasible) full maximum likelihood estimator.

The time domain Gaussian QML estimator is based on treating the $\{X_t\}$ as if they were Gaussian, even though in general they will not be Gaussian. Then we can write -2 times the log likelihood function as

$$L(\theta) = \log|\Sigma_{x,\theta}| + (x-\mu_x)'\Sigma_{x,\theta}^{-1}(x-\mu_x) \tag{4}$$

where $x = (x_1, \ldots, x_n)'$, θ denotes the parameter vector (consisting of the parameters in the model for $\{Y_t\}$ together with σ_η^2), and μ_x, $\Sigma_{x,\theta}$ are, respectively, the expected value of x and the covariance matrix for x under the model θ. Deo (1995) has established the $\sqrt{n}$-consistency and asymptotic normality of the time domain Gaussian QML estimator. Beyond this theoretical result, there are few if any empirical results available on the performance of this estimator, largely due to computational obstacles, i.e., the calculation of the entries of $\Sigma_{x,\theta}$, the determinant $|\Sigma_{x,\theta}|$ and the quadratic form $(x-\mu_x)'\Sigma_{x,\theta}^{-1}(x-\mu_x)$. These obstacles can be surmounted, however.

In fact, $L(\theta)$ may be calculated in $O(n\log^{3/2} n)$ operations, in the case where $\{Y_t\}$ is assumed to obey an $ARFIMA(p,d,q)$ model. This is achieved by using the Fast Fourier Transform (FFT), which is readily available in standard software such as Matlab and S–Plus. We briefly sketch the approach, described in detail in Chen, Hurvich and Lu (2006). Since $\{X_t\}$ is weakly stationary, the entries of $\Sigma_{x,\theta}$ are constant along the diagonals, i.e., $\Sigma_{x,\theta}$ is a Toeplitz matrix. The entire matrix is determined by the autocovariances of $\{X_t\}$ at lags $0, \ldots, n-1$. However, it is important here to avoid actually computing the full matrix $\Sigma_{x,\theta}$ since this would require at least n^2 operations, resulting in extremely slow performance when n is large, say, in the hundreds or thousands. Since the autocovariances of $\{X_t\}$ are identical to those of $\{Y_t\}$, calculation of the entries of $\Sigma_{x,\theta}$ reduces essentially to the calculation of the autocovariances of an $ARFIMA(p,d,q)$ model. Analytical expressions for these autocovariances were obtained by Sowell (1992). These expressions involve the hypergeometric function. Numerically, the autocovariances can be computed to any desired degree of accuracy in $O(n\log n)$ operations using the algorithm of Bertelli and Caporin (2002). Chen, Hurvich and Lu (2006) present a preconditioned conjugate gradient (PCG) algorithm for computing the quadratic form in (4) in $O(n\log^{3/2} n)$ operations. They also present an accurate approximation to the determinant term in (4) due to Böttcher and

Silbermann (1999) which can be computed in $O(1)$ operations. The PCG method for calculating the likelihood is faster than the $O(n^2)$ that would be required based on the algorithm of Levinson (1946) as advocated by Sowell (1992).

Breidt, Crato and de Lima (1998) proposed to estimate the parameters of the LMSV model from $\{X_t\}$ using the Whittle approximation to the likelihood function. Given data $x_1, \ldots, x_n$, define the periodogram

$$I_j = |\sum_{t=1}^{n} x_t \exp(-i\omega_j t)|^2/(2\pi n) \;\; j = 1, \ldots, n-1 \;\; ,$$

where $\omega_j = 2\pi j/n$ are the Fourier frequencies. Mean correction in the definition above is not necessary since it would not change the values of I_j for $j > 0$. The Whittle approximation for $-2\log$ likelihood is

$$L_W(\theta) = \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} \{\log f_{X,\theta}(\omega_j) + I_j/f_{X,\theta}(\omega_j)\}$$

where $f_{X,\theta}(\omega_j)$ is the spectral density for X at frequency (ω_j) under the model θ. It is easy to compute $L_W(\theta)$ since $\{I_j\}$ can be computed in $O(n \log n)$ operations using the FFT, and since $f_{X,\theta}$ is the sum of a constant and an $ARFIMA$ spectral density, which has a simple analytical form. Breidt, Crato and de Lima established the consistency of the Whittle estimator. Hosoya (1997) presents results on the $\sqrt{n}$-consistency and asymptotic normality of the Whittle estimator. Perez and Ruiz (2001) have studied the empirical properties of the Whittle estimator for LMSV models.

4 Semiparametric Estimation

In a preliminary econometric analysis, it is often of interest to try to gauge the existence and strength of long memory without imposing a fully parametric model. An easily implemented semiparametric estimator of d is the log-periodogram regression estimator $\hat{d}_{GPH}$ of Geweke and Porter-Hudak (1983), obtained as $-1/2$ times the least-squares slope estimate in a linear regression of $\{\log I_j\}_{j=1}^{m}$ on $\{\log|1 - \exp(-i\omega_j)|\}_{j=1}^{m}$, where m tends to ∞ more slowly than n. The $\sqrt{m}$-consistency and asymptotic normality of $\hat{d}_{GPH}$ assuming Gaussianity were obtained by Robinson (1995a) with trimming of low frequencies, and by Hurvich, Deo and Brodsky (1998) without trimming. The latter paper also showed that under suitable regularity conditions the optimal choice of m, minimizing the asymptotic mean squared error, is of order $n^{4/5}$. The regularity conditions were imposed on the short-memory component of the spectral density. For any weakly stationary process with memory

parameter d and spectral density f, the short-memory component is defined by $f^*(\omega) = |1 - \exp(-i\omega)|^{2d} f(\omega)$. The results described above do not apply directly to the estimator $\hat{d}_{GPH}$ based on the log squared returns $\{X_t\}$ in the LMSV model, since in general $\{X_t\}$ will be non-Gaussian (and nonlinear).

For the LMSV model, Deo and Hurvich (2001) established the $\sqrt{m}$-consistency and asymptotic normality of $\hat{d}_{GPH}$ based on $\{X_t\}$ under suitable smoothness conditions on the short-memory part of the spectral density of the signal $\{Y_t\}$. Under these conditions the short-memory part of the spectral density of the log squared returns $\{X_t\}$ behaves like $C + \omega^\beta$ as $\omega \to 0^+$ where $C > 0$ and $\beta = 2d \in (0, 1)$. The resulting MSE-optimal choice for m is of order $n^{2\beta/(2\beta+1)}$ and the corresponding mean squared error of $\hat{d}_{GPH}$ is of order $n^{-2\beta/(2\beta+1)}$. Thus, in the LMSV case the optimal rate of convergence of the mean squared error of $\hat{d}_{GPH}$ depends on d and becomes slower as d decreases. This is due to the presence of the noise term in (3) which induces a negative bias in $\hat{d}_{GPH}$. For a given value of d, the bias becomes more severe as larger values of m are used. Even for d close to 0.5, this bias is still problematic as the optimal rate of convergence becomes of order $n^{-2/3}$, much slower than the $O(n^{-4/5})$ rate attained in the Gaussian case, under suitable smoothness conditions.

Hurvich and Ray (2003) introduced a local Whittle estimator of d based on log squared returns in the LMSV model. Hurvich, Moulines and Soulier (2005) established theoretical properties of this semiparametric estimator $\hat{d}_{LW}$, which is a generalization of the Gaussian semiparametric estimator $\hat{d}_{GSE}$ (Künsch (1987); Robinson (1995b). The results of Arteche (2004) imply that in the LMSV context the GSE estimator suffers from a similar limitation as the GPH estimator: in order to attain $\sqrt{m}$-consistency and asymptotic normality the bandwidth m in $\hat{d}_{GSE}$ cannot approach ∞ faster than $n^{2\beta/(2\beta+1)}$, where $\beta = 2d$. The local Whittle estimator avoids this problem by directly accounting for the noise term in (3). From (3), it follows that as $\omega \to 0^+$ the spectral density of the log squared returns behaves as

$$f_X(\omega) \sim G\omega^{-2d}(1 + h(d, \theta, \omega))$$

where $G = f_Y^*(0)$, $f_Y^*(\omega) = |\omega|^{2d} f_Y(\omega)$, $h(d, \theta, \omega) = \theta\omega^{2d}$, and $\theta = \sigma_\eta^2/\{2\pi f_Y^*(0)\}$. We assume here (as did Deo and Hurvich (2001) as well as Hurvich, Deo and Brodsky (1998)) that f_Y^* satisfies a local Lipschitz condition of order 2, as would be the case if $\{Y_t\}$ is a stationary invertible $ARFIMA$ or fractional Gaussian noise process.

The local Whittle contrast function, based on the observations $x_1, \ldots, x_n$, is defined as

$$\hat{W}_m(\tilde{d}, \tilde{G}, \tilde{\theta}) = \sum_{j=1}^{m} \left\{ \log\left(\tilde{G}\omega_j^{-2\tilde{d}}(1 + h(\tilde{d}, \tilde{\theta}, \omega_j)\right) + \frac{I_j}{\tilde{G}\omega_j^{-2\tilde{d}}(1 + h(\tilde{d}, \tilde{\theta}, \omega_j))} \right\}.$$

Concentrating $\tilde{G}$ out of $\hat{W}_m$ yields the profile likelihood

$$\hat{J}_m(\tilde{d},\tilde{\theta}) = \log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{\omega_j^{2\tilde{d}} I_j}{1+h(\tilde{d},\tilde{\theta},\omega_j)}\right) + m^{-1}\sum_{j=1}^{m}\log\{\omega_j^{-2\tilde{d}}(1+h(\tilde{d},\tilde{\theta},\omega_j))\}.$$

The local Whittle estimator is any minimizer of the empirical contrast function $\hat{J}_m$ over the admissible set $\mathcal{D}_n \times \Theta_n$ (which may depend on the sample size n):

$$(\hat{d}_{LW},\hat{\theta}_{LW}) = \arg\min_{(\tilde{d},\tilde{\theta})\in\mathcal{D}_n\times\Theta_n} \hat{J}_m(\tilde{d},\tilde{\theta}).$$

Under suitable regularity conditions, Hurvich, Moulines and Soulier (2005) show that if $m \to \infty$ faster than $n^{4d/(4d+1)+\delta}$ for some arbitrarily small $\delta > 0$ and if $m^5/n^4 \log^2 m \to 0$, then $m^{1/2}(\hat{d}_{LW} - d)$ is asymptotically Gaussian with zero mean and variance $(1+d)^2/(16\,d^2)$. The first condition on m above is a lower bound, implying that the m for $\hat{d}_{LW}$ must increase faster than the *upper* bound on m needed for $\sqrt{m}(\hat{d}_{GPH} - d)$ to be asymptotically Gaussian with zero mean. Nevertheless, if we allow m to increase sufficiently quickly, the estimator $\hat{d}_{LW}$ attains the rate (to within a logarithmic term) of $O_p(\sqrt{n^{-4/5}})$, essentially the same rate as attained by $\hat{d}_{GPH}$ in the Gaussian case and much faster than the rate attained by either $\hat{d}_{GPH}$ or $\hat{d}_{GSE}$ in the LMSV case.

Accurate finite-sample approximations to the variance of $\hat{d}_{LW}$ are given in Hurvich and Ray (2003).

Sun and Phillips (2003) proposed a nonlinear log-periodogram regression estimator $\hat{d}_{NLP}$ of d, using Fourier frequencies $\omega_1, \ldots, \omega_m$. They assumed a model of form (3) in which the signal is a Gaussian long memory process and the noise is a Gaussian white noise. This rules out most LMSV models, since $\log e_t^2$ is typically non-Gaussian. They partially account for the noise term $\{\eta_t\}$ in (3), through a first-order Taylor expansion about zero of the spectral density of the observations. They establish the asymptotic normality of $m^{1/2}(\hat{d}_{NLP} - d)$ under assumptions including $n^{-4d}m^{4d+1/2} \to \text{Const}$. Thus, $\hat{d}_{NLP}$, with a variance of order $n^{-4d/(4d+1/2)}$, converges faster than the GPH estimator, but unfortunately still arbitrarily slowly if d is sufficiently close to zero.

Beyond estimation of d, a related problem of interest is semiparametric testing of the null hypothesis $d = 0$ in the LMSV model, i.e., testing for long memory in volatility. Most existing papers on LMSV models make use of the assumption that $d > 0$ so the justification of the hypothesis test requires additional work. The ordinary t-test based on either $\hat{d}_{GPH}$ or $\hat{d}_{GSE}$ was justified in Hurvich and Soulier (2002) and Hurvich, Moulines and Soulier (2005), respectively, without strong restrictions on the bandwidth.

5 Generalizations of the LMSV Model

It is possible to relax the assumption that $\{Y_t\}$ and $\{e_t\}$ are independent in (1). A contemporaneous correlation between $\{e_t\}$ and the shocks in the model for $\{Y_t\}$ was allowed for in Hurvich, Moulines and Soulier (2005), as well as Hurvich and Ray (2003), Surgailis and Viano (2002). See Hurvich and Ray (2003) for more details on estimating the leverage effect, known in the (exponential) GARCH models, where the sign of the return in period t affects the conditional variance for period $t+1$.

It is possible to replace the Gaussianity assumption for $\{Y_t\}$ in the LMSV model by a linearity assumption. This was done in Hurvich, Moulines and Soulier (2005) and Arteche (2004), among others. Surgailis and Viano (2002) showed that under linearity for $\{Y_t\}$ and other weak assumptions, powers of the absolute returns have long memory, with the same memory parameter as $\{Y_t\}$. This result does not require any assumption about the dependence between $\{Y_t\}$ and $\{e_t\}$.

It is also possible to relax the assumption that $d < 1/2$ in the LMSV model. If $d \in (1/2, 1)$ we can say that the volatility is mean reverting but not stationary. Hurvich, Moulines and Soulier (2005) proved consistency of $\hat{d}_{LW}$ for $d \in (0, 1)$ and proved the $\sqrt{m}$-consistency and asymptotic normality of $\hat{d}_{LW}$ for $d \in (0, 3/4)$.

6 Applications of the LMSV Model

We briefly mention some applications of the LMSV and related models. Deo, Hurvich and Lu (2006) consider using the (parametric) LMSV model to construct forecasts of realized volatility. The forecast is given as a linear combination of present and past squared returns. The forecast weights are obtained using the PCG algorithm.

A long memory stochastic duration (LMSD) model was introduced in Deo, Hsieh and Hurvich (2005) to describe the waiting times (durations) between trades of a financial asset. The LMSD model has the same mathematical form as the LMSV model, except that the multiplying shocks have a distribution with positive support.

Smoothing of the volatility in LMSV models was considered by Harvey (1998), who gave a formula for the minimum mean squared error linear estimator (MMSLE) of $\{Y_t\}_{t=1}^n$ based on the observations $\{X_t\}_{t=1}^n$. Computation of the coefficients in the linear combination involves the solution of a Toeplitz system, and the MMSLE can be efficiently computed using the PCG algorithm. Nevertheless, the MMSLE methodology suffers from some drawbacks, as described (in the LMSD context) in Deo, Hsieh and Hurvich (2005).

References

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2001): Modeling and forecasting realized volatility. *Journal of the American Statistical Association* **96**, 42–55.

Arteche, J. (2004): Gaussian semiparametric estimation in long memory in stochastic volatility and signal plus noise models. *Journal of Econometrics* **119**, 131–154.

Bertelli, S. and Caporin, M. (2002): A note on calculating autocovariances of long-memory processes. *Journal of Time Series Analysis* **23**, 503–508.

Böttcher, A. and Silbermann, B. (1999): *Introduction to Large Truncated Toeplitz Matrices*. Springer Verlag, New York.

Breidt, F. J., Crato, N., and de Lima, P. (1998): The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics* **83**, 325–348.

Chen, W.W., Hurvich, C.M. and Lu, Y. (2006): On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series. *Journal of the American Statistical Association* **101**, 812–822.

Deo, R. (1995): On GMM and QML estimation for the long memory stochastic volatility model. *Working Paper*.

Deo, R.S. and Hurvich, C.M. (2001): On the log periodogram regression estimator of the memory parameter in long memory stochastic volatility models. *Econometric Theory* **17**, 686–710.

Deo, R.S., Hsieh, M. and Hurvich, C.M. (2005): Tracing the source of long memory in volatility. *Working paper*. http://w4.stern.nyu.edu/emplibrary/LMSDdata.pdf

Deo, R.S., Hurvich, C.M. and Lu, Y. (2006): Forecasting realized volatility using a long memory stochastic volatility model: estimation, prediction and seasonal adjustment. *Journal of Econometrics* **131**, 29–58.

Ding, Z., Granger, C. and Engle, R.F. (1993): A long memory property of stock market returns and a new model. *Journal of Empirical Finance* **1**, 83–106.

Geweke, J. and Porter-Hudak, S. (1983): The estimation and application of long memory time series models. *Journal of Time Series Analysis* **4**, 221–238.

Harvey, A.C. (1998): Long memory in stochastic volatility. *In: Knight, J. and Satchell, S. (Eds.): Forecasting Volatility in Financial Markets*, 307–320. Butterworth-Heinemann, London.

Harvey, A.C., Ruiz, E. and Shephard, N. (1994): Multivariate stochastic volatility models. *Review of Economic Studies* **61**, 247–264.

Hosoya, Y. (1997): A limit theory for long-range dependence and statistical inference on related models. *Annals of Statistics* **25**, 105–137.

Hurvich, C.M., Deo, R.S. and Brodsky, J. (1998): The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis* **19**, 19–46.

Hurvich, C.M., Moulines, E. and Soulier, Ph. (2005): Estimating long memory in volatility. *Econometrica* **73**, 1283–1328.

Hurvich, C.M. and Ray, B.K. (2003): The local Whittle estimator of long-memory stochastic volatility. *Journal of Financial Econometrics* **1**, 445–470.

Hurvich, C.M. and Soulier, Ph. (2002): Testing for long memory in volatility. *Econometric Theory* **18**, 1291–1308.

Künsch, H.R. (1987): Statistical aspects of self-similar processes. *In: Proceedings of the World Congress of the Bernoulli Society (Tashkent)* **1**, 67–74.

Levinson, N. (1946): The Wiener RMS (root mean square) error criterion in filter design and prediction. *Journal of Mathematical Physics* **25**, 261–178.

Perez, A. and Ruiz, E. (2001): Finite sample properties of a QML estimator of stochastic volatility models with long memory. *Economics Letters* **70**, 157–164.

Robinson, P.M. (1995a): Log-periodogram regression of time series with long range dependence. *Annals of Statistics* **23**, 1043–1072.

Robinson, P.M. (1995b): Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* **24**, 1630–1661.

Robinson, P.M. (2001): The memory of stochastic volatility models. *Journal of Econometrics* **101**, 195–218.

Robinson, P.M. and Zaffaroni, P. (1997): Modelling nonlinearity and long memory in time series. *Fields Institute Communications* **11**, 161–170.

Robinson, P.M. and Zaffaroni, P. (1998): Nonlinear time series with long memory: a model for stochastic volatility. *Journal of Statistical Planning and Inference* **68**, 359–371.

Sowell, F. (1992): Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models. *Journal of Econometrics* **53**, 165–188.

Surgailis, D. and Viano, M.C. (2002): Long memory properties and covariance structure of the EGARCH model. *ESAIM: Probability and Statistics* **6**, 311–329.

Sun, Y. and Phillips, P.C.B. (2003): Nonlinear log-periodogram regression for perturbed fractional processes. *Journal of Econometrics* **115**, 355–389.

Taqqu, M.S. (2003): Fractional Brownian motion and long-range dependence *In: Doukhan, P., Oppenheim, G. and Taqqu, M.S. (Eds.): Theory and Applications of Long-Range Dependence*, 5–38. Birkhäuser, Boston.

Taylor, S.J. (1986): *Modelling Financial Time Series*. Wiley, New York.

Extremes of Stochastic Volatility Models

Richard A. Davis and Thomas Mikosch

Abstract We consider extreme value theory for stochastic volatility processes in both cases of light-tailed and heavy-tailed noise. First, the asymptotic behavior of the tails of the marginal distribution is described for the two cases when the noise distribution is Gaussian or heavy-tailed. The sequence of point processes, based on the locations of the suitable normalized observations from a stochastic volatility process, converges in distribution to a Poisson process. From the point process convergence, a variety of limit results for extremes can be derived. Of special note, there is no extremal clustering for stochastic volatility processes in both the light- and heavy-tailed cases. This property is in sharp contrast with GARCH processes which exhibit extremal clustering (i.e., large values of the process come in clusters).

1 Introduction

The simple stochastic volatility process $(X_t)_{t\in\mathbb{Z}}$ is given by the equation

$$X_t = \sigma_t Z_t\,, \quad t \in \mathbb{Z}\,, \tag{1}$$

where (Z_t) is iid, $(\sigma_t)_{t\in\mathbb{Z}}$ is the log-linear Gaussian process given by

Richard A. Davis
Department of Statistics, 1255 Amsterdam Avenue, Columbia University, New York, NY 10027, U.S.A., www.stat.columbia.edu/∼rdavis, e-mail: rdavis@stat.columbia.edu

Thomas Mikosch
Laboratory of Actuarial Mathematics, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark, www.math.ku.dk/∼mikosch, e-mail: mikosch@math.ku.dk

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_15,

$$2\log\sigma_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}\,,$$

with $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, and the sequence (η_t) is iid $\mathrm{N}(0,\tau^2)$ and independent of (Z_t). If $\mathrm{var}(Z_t) < \infty$, then it is customary to assume that (Z_t) is iid with mean 0 and variance 1. In this article, we describe the limiting behavior of the sample maxima,

$$M_n = \max(X_1, \ldots, X_n)\,,$$

of the strictly stationary stochastic volatility sequence (X_t) in the cases that the noise (Z_t) has either a light- or heavy-tailed distribution.

In Section 2, we describe the tail behavior of the marginal distribution of X_1. Point process convergence based on the normalized process is described in Section 3. This provides the key result from which limiting behavior of the extremes of (X_t) can be determined.

Interestingly, and unlike the situation for GARCH processes (see Davis and Mikosch (2008a)), there is no extremal clustering for stochastic volatility processes in both the light- and heavy-tailed cases. That is, large values of the processes do not come in clusters. More precisely, the large sample behavior of M_n is the same as that of the maxima of the associated iid sequence $(\widehat{X}_t)$, where $\widehat{X} \stackrel{d}{=} X$.

2 The Tail Behavior of the Marginal Distribution

2.1 The light-tailed case

For the model given by (1), assume further that the noise (Z_t) is iid N(0,1). Notice that the log of the squares of the process, i.e.,

$$Y_t = \log X_t^2 \tag{2}$$

is the superposition of a linear Gaussian process with iid log-χ_1^2 distributed noise. Since $\log Z_t^2$ is distributed as the log of a χ_1^2 random variable, its cumulant generating function is given by

$$\begin{aligned}\log E\exp\left\{\lambda \log \chi_1^2\right\} &= \lambda\log 2 + \log\Gamma(1/2+\lambda) - \log\Gamma(1/2)\\ &= \lambda\log\lambda + \lambda(\log 2 - 1) + (\log 2)/2 + h_0(\lambda)\,,\end{aligned}$$

where the remainder $h_0(\lambda) = O(1/\lambda)$ as $\lambda \to \infty$. The cumulant generating function of Y_t is

$$\kappa(\lambda) = \log E\mathrm{e}^{\lambda Y_t} = \lambda^2\widetilde{\sigma}^2/2 + \lambda\log 2 + \log\Gamma(1/2+\lambda) - \log\Gamma(1/2)$$
$$= \lambda^2\widetilde{\sigma}^2/2 + \lambda\log\lambda + \lambda(\log 2 - 1) + (\log 2)/2 + O(1/\lambda),$$

where

$$\widetilde{\sigma}^2 = \mathrm{var}\left(\sum_{j=0}^{\infty}\psi_j\eta_{t-j}\right) = \tau^2\sum_{j=0}^{\infty}\psi_j^2\,. \tag{3}$$

The following proposition, due to Breidt and Davis (1998), describes the tail behavior of Y_t. The proof of this result is based on the asymptotic normality of the Esscher transform of the distribution of Y which can be viewed as the saddlepoint approximation to the cumulant generating function, see Jensen (1995). This technique was also used by Feigin and Yashchin (1983) and Davis and Resnick (1991).

Proposition 1 *For the log-squared volatility process* (Y_t) *defined in* (2), *we have*

$$P(Y_t > x) \sim \frac{\widetilde{\sigma}^2}{\sqrt{\pi}}\exp\left\{-\frac{x^2}{2\widetilde{\sigma}^2} + \frac{x\log x}{\widetilde{\sigma}^2} + \frac{(k-1)x}{\widetilde{\sigma}^2} - \frac{(k+\widetilde{\sigma}^2)\log x}{\widetilde{\sigma}^2} - \frac{\log^2 x}{2\widetilde{\sigma}^2}\right.$$
$$\left. -\frac{k^2}{2\widetilde{\sigma}^2} + O\left(\frac{\log^2 x}{x}\right)\right\},$$

as $x\to\infty$, *where* $k = \log(2/\widetilde{\sigma}^2)$.

By symmetry of X_t,

$$P(X_t > x) = \frac{1}{2}P(|X_t| > x) = \frac{1}{2}P(\log X_t^2 > 2\log x)\,,$$

and then the asymptotic behavior of $P(X_t > x)$ as $x\to\infty$ is straightforward from Proposition 1.

2.2 The heavy-tailed case

Now assume that Z_t has regularly varying tail probabilities with index $\alpha > 0$. This means that the distribution of $|Z_t|$ is regularly varying with index α, i.e.,

$$P(|Z_t| > x) = L(x)\,x^{-\alpha} \tag{4}$$

as $x\to\infty$, where $L(\cdot)$ is a slowly varying function at infinity (see Section 4 of Davis and Mikosch (2008a)), and that the tail balancing condition,

$$\lim_{x\to\infty} \frac{P(Z_t > x)}{P(|Z_t| > x)} = p \quad \text{and} \quad \lim_{x\to\infty} \frac{P(Z_t \le -x)}{P(|Z_t| > x)} = q\,, \tag{5}$$

where $p + q = 1$ for some $p \in [0, 1]$, holds. Then, by virtue of Breiman's result (see equation (16) in Davis and Mikosch (2008a)), the distribution of X_t inherits the same tail behavior as Z_t.

Proposition 2 *Under the regularly varying and tail-balancing assumptions* (4) *and* (5)*, we have*

$$P(X_t > x) \sim E(\sigma_t^\alpha)\, P(Z_t > x) \quad \textit{and} \quad P(X_t \le -x) \sim E(\sigma_t^\alpha)\, P(Z_t \le -x)\,,$$

as $x \to \infty$.

Proposition 2 remains valid even if $\log \sigma_t$ is not a linear Gaussian process. In order to apply Breiman's result, one only needs that σ_t is independent of Z_t and $E(\sigma_t^{\alpha+\epsilon}) < \infty$ for some $\epsilon > 0$.

3 Point Process Convergence

3.1 Background

The theory of point processes plays a central role in extreme value theory. For example, the limiting distribution of order statistics, such as the kth largest, is often easy to derive from the convergence of a particular sequence of point processes. To illustrate this notion, suppose $(\widehat{X}_t)$ is an iid sequence of random variables with the same common distribution function F as X_t. Further assume that there exist sequences of constants $a_n > 0$ and b_n such that

$$P(a_n^{-1}(\widehat{M}_n - b_n) \le x) = F^n(a_n x + b_n) \to G(x) \tag{6}$$

for all x, where $\widehat{M}_n = \max(\widehat{X}_1, \ldots, \widehat{X}_n)$ and G is a nondegenerate distribution function. Then by the extremal types theorem, G has to be an extreme value distribution of which there are only three types, see Leadbetter et al. (1983). Moreover, by taking logarithms and using a Taylor series expansion, (6) holds if and only if for any $x \in \mathbb{R}$,

$$n\,(1 - F(a_n x + b_n)) \to -\log G(x)\,.$$

or,[1] equivalently, if for any $x \in \mathbb{R}$,

$$n\,P(a_n^{-1}(\widehat{X}_1 - b_n) > x) \to -\log G(x)\,. \tag{7}$$

[1] If $G(x) = 0$ we interpret $-\log G(x)$ as ∞.

Now (7) can be strengthened to the statement,

$$n\,P(a_n^{-1}(\widehat{X}_1 - b_n) \in B) \to \nu(B) \tag{8}$$

for all suitably chosen Borel sets B, where the measure ν is defined by its value on intervals of the form $(a, b]$ as

$$\nu(a, b] = \log G(b) - \log G(a)\,. \tag{9}$$

The convergence in (8) can be connected with the convergence in distribution of a sequence of point processes. For a bounded Borel set B in the product space $(0, \infty) \times \mathbb{R}$, define the sequence of point processes $(\widehat{N}_n)$ by

$$\widehat{N}_n(B) = \#\{(j/n, a_n^{-1}(\widehat{X}_j - b_n)) \in B\,, j = 1, 2, \ldots\}\,.$$

If B is the rectangle $(a, b] \times (c, d]$ with $0 \le a < b < \infty$ and $-\infty < c < d < \infty$, then since the $\widehat{X}_j$ are iid, $\widehat{N}_n(B)$ has a binomial distribution with number of trials $[nb] - [na]$ ($[s]$ = integer part of s), and probability of success

$$p_n = P(a_n^{-1}(\widehat{X}_1 - b_n) \in (c, d])\,.$$

Provided $\nu(c, d] < \infty$, it follows from (8) that $\widehat{N}_n(B)$ converges in distribution to a Poisson random variable $N(B)$ with mean $\mu(B) = (b-a)\,\nu(c, d]$. In fact, we have the stronger point process convergence,

$$\widehat{N}_n \xrightarrow{d} N\,, \tag{10}$$

where N is a Poisson process on $(0, \infty) \times \mathbb{R}$ with mean measure $\mu(dt, dx) = dt \times \nu(dx)$ and $\xrightarrow{d}$ denotes convergence in distribution of point processes. For our purposes, $\xrightarrow{d}$ for point processes means that for any collection of *bounded*[2] Borel sets $B_1, \ldots, B_k$ for which $P(N(\partial B_j) > 0) = 0$, $j = 1, \ldots, k$, we have

$$(\widehat{N}_n(B_1), \ldots, \widehat{N}_n(B_k)) \xrightarrow{d} (N(B_1), \ldots, N(B_k))$$

on $\mathbb{R}^k$, see Embrechts et al. (1997), Leadbetter et al. (1983), Resnick (1987).

As an application of (10), define $\widehat{M}_{n,2}$ to be the second largest among $\widehat{X}_1, \ldots, \widehat{X}_n$. Since the event $\{a_n^{-1}(\widehat{M}_{n,2} - b_n) \le y\}$ is the same as $\{\widehat{N}_n((0, 1] \times (y, \infty)) \le 1\}$, we conclude from (10) that

[2] In some cases, especially the heavy-tailed case, the state space of the point process is often defined to be $(0, \infty) \times ([-\infty, \infty] \setminus \{0\})$. On the second product space, the roles of zero and infinity have been interchanged so that bounded sets are now those sets which are bounded away from 0. With this convention, a bounded set on the product space is contained in the rectangle $[0, c] \times ([-\infty, -d] \cup [d, \infty])$ for some positive and finite constants c and d. Under this topology, the intensity measure for the Poisson process defined in Theorem 2 is ensured to be finite on all bounded Borel sets.

$$\begin{aligned} P(a_n^{-1}(\widehat{M}_{n,2} - b_n) \le y) &= P(\widehat{N}_n((0,1] \times (y,\infty)) \le 1) \\ &\to P(N((0,1] \times (y,\infty)) \le 1)) \\ &= G(y)\,(1 - \log G(y))\,. \end{aligned}$$

Similarly, the joint limiting distribution of $(\widehat{M}_n, \widehat{M}_{n,2})$ can be calculated by noting that for $y \le x$, $\{a_n^{-1}(\widehat{M}_n - b_n) \le x, a_n^{-1}(\widehat{M}_{n,2} - b_n) \le y\} = \{\widehat{N}_n((0,1] \times (x,\infty)) = 0, \widehat{N}_n((0,1] \times (y,x]) \le 1\}$. Hence,

$$\begin{aligned} &P(a_n^{-1}(\widehat{M}_n - b_n) \le x, a_n^{-1}(\widehat{M}_{n,2} - b_n) \le y) \\ &= P(\widehat{N}_n((0,1] \times (x,\infty)) = 0, \widehat{N}_n((0,1] \times (y,x]) \le 1) \\ &\to P(N((0,1] \times (x,\infty)) = 0, N((0,1] \times (y,x]) \le 1)) \\ &= G(y)(1 + \log G(x) - \log G(y))\,. \end{aligned}$$

3.2 Application to stochastic volatility models

The point process convergence in (10) can be extended to general stationary time series provided a mixing condition and a local dependence condition (such as D and D' in Leadbetter et al. (1983)) hold. The mixing condition governs how fast a certain class of events become independent as their time separation increases. Typically, many time series models, including stochastic volatility processes, satisfy a mixing condition such as strong mixing. (For stochastic volatility processes, see the discussion for strong mixing given in Section 2 of Davis and Mikosch (2008a).) On the other hand, the dependence condition D' restricts the clustering of extremes. That is, given an observation at time t is large, the probability that any of its neighboring observations are also large is quite small. The stochastic volatility processes (X_t) given in (1) with either light- or heavy-tailed noise satisfies generalized versions of conditions D and D'; see Breidt and Davis (1998), Davis and Mikosch (2001). Thus the point process convergence in (10) holds. This result is recorded in the following two theorems whose proofs can be found in Breidt and Davis (1998) for the light-tailed case (Theorem 1) and in Davis and Mikosch (2001) for the heavy-tailed case (Theorem 2).

3.2.1 The light-tailed case

Theorem 1 *Suppose (X_t) is the stochastic volatility process defined in* (1)*, where the noise (Z_t) is iid* $\mathrm{N}(0,1)$ *and the autocorrelation function*

$$\rho(h) = \text{corr}(\log \sigma_t^2, \log \sigma_{t+h}^2)$$

decays at the rate $\rho(h) = o(1/\log h)$ *as* $h \to \infty$. *Let the constants* a_n *and* b_n *be defined by*

$$a_n = \widetilde{\sigma}\ (2\log n)^{-1/2} = \widetilde{\sigma}/(\sqrt{2}\, d_n) \tag{11}$$

where $d_n = (\log n)^{1/2}$, $\widetilde{\sigma}^2$ *is given in* (3), *and*

$$b_n = c_1 d_n + \log d_n + c_2 + c_3 \frac{\log d_n}{d_n} + c_4 \frac{1}{d_n}, \tag{12}$$

where

$$c_1 = \left(2\widetilde{\sigma}^2\right)^{1/2},\ c_2 = \frac{3}{2}\log 2 - \frac{1}{2}\log \widetilde{\sigma}^2 - 1,\ c_3 = -\frac{\widetilde{\sigma}}{\sqrt{2}},$$

and

$$c_4 = -\frac{1}{2(2\widetilde{\sigma}^2)^{1/2}}\left(1 + \widetilde{\sigma}^2 \log(2\pi)\right).$$

Then, with $Y_t = \log X_t^2$, *the limit in* (8) *holds with* $\widehat{X}_1$ *replaced with* Y_1 *and* $G(x) = \exp\{-\exp\{-x\}\}$ *in* (9). *Moreover,* $N_n \xrightarrow{d} N$, *where* N_n *is the point process defined by*

$$N_n(B) = \#\{(j/n, a_n^{-1}(Y_j - b_n)) \in B,\ j = 1, 2, \ldots\},$$

and N *is a Poisson point process on* $(0,\infty)\times(-\infty,\infty)$ *with intensity measure* $dt \times \nu(dx)$.

The theorem shows that for a wide class of stochastic volatility models driven with normal noise, the extremes of (Y_t) can be normalized independently of the covariance structure in $(\log \sigma_t^2)$, and the same limiting distribution is obtained in all cases. In finite samples, however, the degree of dependence in this linear process does affect the goodness-of-fit of the limiting distribution (see Figure 1 of Breidt and Davis (1998)).

Defining the maximum of the log-squared volatility sequence by $M_n^Y = \max(Y_1, \ldots, Y_n)$, the limit distribution of the maxima can be determined directly from the theorem in the way explained in Section 3.1 and is given by

$$\begin{aligned} P(a_n^{-1}(M_n^Y - b_n) \le x) &= P(N_n((0,1]\times(x,\infty)) = 0) \\ &\to P(N((0,1]\times(x,\infty)) = 0) \\ &= \mathrm{e}^{-\mathrm{e}^{-x}}, \quad x \in \mathbb{R}. \end{aligned} \tag{13}$$

The limit is the Gumbel distribution. It is one of the extreme value distributions, see Leadbetter et al. (1983).

The limiting distribution for the maxima $M_n^{|X|} = \max(|X_1|, \ldots, |X_n|)$ of the absolute values of the original (untransformed) volatility process (X_t) can also be found from (13). Indeed, observe that for any $x \in \mathbb{R}$,

$$
\begin{aligned}
&P(a_n^{-1}(M_n^Y - b_n) \le x) \\
&= P(M_n^{|X|} \le \mathrm{e}^{\,0.5\,(a_n\,x + b_n)}) \\
&= P(\mathrm{e}^{\,-0.5\,b_n}\,(0.5\,a_n)^{-1}\,(M_n^{|X|} - \mathrm{e}^{\,0.5\,b_n}) \le\, x + o(1))\,,
\end{aligned}
$$

where we used the Taylor expansion argument $\exp\{0.5\,a_n\,x\} = 1 + 0.5\,a_n\,x + o(a_n)$. Another Taylor expansion yields as $n \to \infty$

$$
\mathrm{e}^{\,0.5\,b_n}\,0.5\,a_n \sim 2^{-3/4}\,\widetilde{\sigma}^{-3/2}\,\mathrm{e}^{\,-0.5}\,(\log d_n)^{1/2}\mathrm{e}^{\,\widetilde{\sigma}\,d_n/\sqrt{2}} = \widetilde{a}_n\,. \tag{14}
$$

Combining the arguments above and recalling that the Gumbel distribution is continuous, we may conclude the following.

Corollary 1 *Under the conditions of Theorem 1,*

$$
P(\widetilde{a}_n^{-1}\,(M_n^{|X|} - \mathrm{e}^{\,0.5\,b_n}) \le x) \to \mathrm{e}^{\,-\mathrm{e}^{\,-x}}\,, \quad x \in \mathbb{R}\,,
$$

where $\widetilde{a}_n$ *and* b_n *are defined in* (14) *and* (12)*, respectively.*

One can also recover the limit distribution for the maxima $M_n^X = \max(X_1, \ldots, X_n)$ of the original series. The proof, which we sketch here, follows the argument given in Haan et al. (1989) as adapted by Breidt and Davis (1998). First note that $X_t = |X_t|\,r_t$, where $(r_t) = (\mathrm{sign}(X_t))$ is an iid sequence with $P(r_t = 1) = P(r_t = -1) = 0.5$ and independent of $(|X_t|)$. For x fixed, set $B_n = N_n((0,1] \times (x, \infty))$, $u_n = a_n x + b_n$, and $v_n^2 = \exp\{u_n\}$. If $1 \le \tau_1 < \tau_2 < \cdots$ denote the times at which (X_t^2) exceeds v_n^2, then

$$
\begin{aligned}
P(M_n^X \le v_n) &= \sum_{k=0}^{\infty} P\left(B_n = k, M_n^X \le v_n\right) \\
&= \sum_{k=0}^{\infty} P\left(B_n = k, r_{\tau_1} = -1, \ldots, r_{\tau_k} = -1\right), \qquad (15)
\end{aligned}
$$

because the event $\{B_n = k, M_n^X \le v_n\}$ corresponds to the event that there are exactly k exceedances of v_n^2 by $X_1^2, \ldots, X_n^2$, where each such exceedance corresponds to a negative sign of the respective noise term. Since B_n is independent of the signs of the X_t and the random times τ_i depend only on $|X_{\tau_i}|$ and are independent of (r_{τ_i}), the right-hand side of (15) is equal to

$$
\begin{aligned}
\sum_{k=0}^{\infty} P\left(B_n = k\right) 2^{-k} &\to \sum_{k=0}^{\infty} P\left(N((0,1] \times (x, \infty)) = k\right) 2^{-k} \\
&= \sum_{k=0}^{\infty} \frac{(\mathrm{e}^{\,-x}/2)^k\,\mathrm{e}^{\,-\mathrm{e}^{\,-x}}}{k!} \\
&= \mathrm{e}^{\,-\mathrm{e}^{\,-x}/2}\,,
\end{aligned}
$$

where the first equality follows from dominated convergence. Using a Taylor series expansion on v_n as above, we obtain the following limit result for M_n^X.

Corollary 2 *Under the conditions of Theorem 1,*

$$P(\widetilde{a}_n^{-1}\,(M_n^X - \mathrm{e}^{\,0.5\,b_n}) \le x) \to \mathrm{e}^{\,-\mathrm{e}^{\,-x}}\,, \quad x \in \mathbb{R}\,,$$

where $\widetilde{a}_n$ *and* b_n *are defined in* (14) *and* (12)*, respectively.*

3.2.2 The heavy-tailed case

Theorem 2 *Suppose* (X_t) *is the stochastic volatility process given by (1), where* Z_t *satisfies (4) and (5). Let* a_n *be the* $(1 - n^{-1})$*-quantile of* $|X_t|$*, i.e.,* $a_n = \inf\{x : P(|X_t| > x) \le n^{-1}\}$ *and define the point process* N_n *by*

$$N_n(B) = \#\{(j/n, a_n^{-1} X_j) \in B,\, j = 1, 2, \ldots\}\,.$$

Then $N_n \xrightarrow{d} N$*, where* N *is a Poisson point process on* $(0,\infty) \times (-\infty,\infty)$ *with intensity measure* $dt \times \nu(dx)$*, and*

$$\nu(dx) = \left(p\,\alpha\,x^{-\alpha-1} 1_{(0,\infty)}(x) + q\,\alpha\,(-x)^{-\alpha-1} 1_{(-\infty,0)}(x)\right) dx\,.$$

Moreover,

$$P(a_n^{-1}\,M_n \le x) \to \mathrm{e}^{\,-p\,x^{-\alpha}}\,, \quad x > 0\,,$$

i.e., the limit is the Fréchet distribution which is one of the extreme value distributions, see Embrechts et al. (1997), Leadbetter et al. (1983), Resnick (1987).

For a stationary process (X_t) that satisfies a general mixing condition, one can often show the existence of a $\theta \in (0, 1]$ such that

$$P(a_n^{-1}(M_n - b_n) \le x) \to G^\theta(x)\,,$$

where the marginal distribution of the process satisfies (7). The parameter θ is called the *extremal index* and measures the level of clustering of extremes for stationary processes. One can interpret $1/\theta$ as the mean cluster size of exceedances above a high threshold. For $\theta = 1$, there is no clustering and so the maxima behave asymptotically the same as the corresponding maxima of the iid sequence with the same marginal distribution. For the stochastic volatility process with either light- or heavy-tailed noise, it follows from Corollary 2 and Theorem 2 that the extremal index is always 1. In contrast, the extremal index for the GARCH process is always less than one; see Davis and Mikosch (2008b). So while both stochastic volatility and GARCH processes exhibit volatility clustering, only the GARCH has clustering of extremes.

References

Breidt, F.J. and Davis, R.A. (1998): Extremes of stochastic volatility models. *Ann. Appl. Probab* **8**, 664–675.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods* (2nd edition). Springer, Berlin, Heidelberg, New York.

Davis, R.A. and Resnick, S.I. (1991): Extremes of moving averages of random variables with finite endpoint. *Ann. Probab.* **19**, 312–328.

Davis, R.A. and Mikosch, T. (2001): Point process convergence of stochastic volatility processes with application to sample autocorrelations. *J. Appl. Probab. Special Volume: A Festschrift for David Vere-Jones* **38A**, 93–104.

Davis, R.A. and Mikosch, T. (2008a): Probabilistic properties of stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 255–267. Springer, New York.

Davis, R.A. and Mikosch, T. (2008b): Extreme value theory for GARCH models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 186–200. Springer, New York.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997): *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

Feigin, P.D. and Yashchin, E. (1983): On a strong Tauberian result. *Z. Wahrsch. Verw. Gebiete* **65**, 35–48.

Haan, L. de, Resnick, S.I., Rootzén, H. and Vries, C.G. de (1989): Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. *Stoch. Proc. Appl.* **32**, 213–224.

Jensen, J.L. (1995): *Saddlepoint Approximations.* Oxford University Press, Oxford.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983): *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

Resnick, S.I. (1987): *Extreme Values, Regular Variation and Point Processes.* Springer, New York.

Multivariate Stochastic Volatility

Siddhartha Chib, Yasuhiro Omori and Manabu Asai

Abstract We provide a detailed summary of the large and vibrant emerging literature that deals with the multivariate modeling of conditional volatility of financial time series within the framework of stochastic volatility. The developments and achievements in this area represent one of the great success stories of financial econometrics. Three broad classes of multivariate stochastic volatility models have emerged: one that is a direct extension of the univariate class of stochastic volatility model, another that is related to the factor models of multivariate analysis and a third that is based on the direct modeling of time-varying correlation matrices via matrix exponential transformations, Wishart processes and other means. We discuss each of the various model formulations, provide connections and differences and show how the models are estimated. Given the interest in this area, further significant developments can be expected, perhaps fostered by the overview and details delineated in this paper, especially in the fitting of high-dimensional models.

Siddhartha Chib
Washington University in St. Louis, Campus Box 1133, 1 Brookings Dr., St. Louis, MO 63130, USA, e-mail: chib@wustl.edu

Yasuhiro Omori
Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan, e-mail: omori@e.u-tokyo.ac.jp

Manabu Asai
Faculty of Economics, Soka University, 1-236 Tangi-cho, Hachioji-shi Tokyo, 192-8577, Japan, e-mail: m-asai@soka.ac.jp

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_16,

1 Introduction

A considerable amount of recent literature on financial econometrics has emerged on the modeling of conditional volatility, spurred by the demand for such models in areas such as portfolio and risk management. Much of the early interest centered on multivariate versions of univariate generalized autoregressive conditional heteroscedasticity (GARCH) models. These generalizations have been ably summarized in recent surveys, for example, those of Bauwens et al. (2006) and Silvennoinen and Teräsvirta (2007). More recently, a large and prolific (parallel) body of literature has developed around generalizations of the univariate stochastic volatility (SV) model. A number of multivariate SV (MSV) models are now available along with clearly articulated estimation recipes. Our goal in this paper is to provide a detailed summary of these various model formulations, along with connections and differences, and discuss how the models are estimated. We aim to show that the developments and achievements in this area represent one of the great success stories of financial econometrics. We note that our treatment does not include any discussion of multivariate modeling of volatility that is relevant for ultra-high-frequency data. Thus, there is no discussion of realized volatility (Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2004)).

To fix notation and set the stage for our developments, the univariate SV model that forms the basis for many MSV models is given by (Ghysels et al. (1996), Broto and Ruiz (2004) and Shephard (2004)):

$$y_t = \exp(h_t/2)\varepsilon_t, \quad t = 1, \ldots, n, \tag{1}$$

$$h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t, \quad t = 1, \ldots, n-1, \tag{2}$$

$$h_1 \sim \mathcal{N}\left(\mu, \sigma_\eta^2/(1-\phi^2)\right), \tag{3}$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} | h_t \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}, \tag{4}$$

where y_t is a univariate outcome, h_t is a univariate latent variable and $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}_m(\mu, \mathbf{\Sigma})$ denote, respectively, a univariate normal distribution with mean μ and variance σ^2, and an m-variate normal distribution with mean vector μ and variance-covariance matrix $\mathbf{\Sigma}$. In this model, conditioned on the parameters $(\mu, \phi, \sigma_\eta^2)$, the first generating equation represents the distribution of y_t conditioned on h_t, and the second generating equation represents the Markov evolution of h_{t+1} given h_t. The conditional mean of y_t is assumed to be zero because that is a reasonable assumption in the setting of high-frequency financial data. The SV model is thus a state-space model, with a linear evolution of the state variable h_t but with a nonlinear measurement equation (because h_t enters the outcome model nonlinearly). Furthermore, from the measurement equation we see that $\mathrm{Var}(y_t|h_t) = \exp(h_t)$, which implies that h_t may be understood as the log of the conditional variance of the outcome. To ensure that the evolution of these log volatilities is

stationarity, one generally assumes that $|\phi| < 1$. Many other versions of the univariate SV model are possible. For example, it is possible to let the model errors have a non-Gaussian fat-tailed distribution, to permit jumps, and incorporate the leverage effect (through a nonzero off-diagonal element in $\mathbf{\Sigma}$). The estimation of the canonical SV model and its various extensions was at one time considered difficult since the likelihood function of these models is not easily calculable. This problem has been fully resolved by the creative use of Monte Carlo methods, primarily Bayesian Markov chain Monte Carlo (MCMC) methods (Jacquier et al. (1994), Kim et al. (1998), Chib et al. (2002) and Omori et al. (2007)). We refer the readers to Asai et al. (2006) for a discussion of how this problem can be addressed in some special cases by non-Bayesian methods. In this survey, on the other hand, we concentrate on Bayesian methods but mention the full range of methods (Bayesian and non-Bayesian) that have been tried for the various models.

In the multivariate case, when one is dealing with a collection of financial time series denoted by $\mathbf{y}_t = (y_{1t}, \ldots, y_{pt})'$, the main goal is to model the time-varying conditional covariance matrix of $\mathbf{y}_t$. There are several ways in which this can be done. A typical starting point is the assumption of series-specific log volatilities h_{tj} $(j \leq p)$ whose joint evolution is governed by a first-order stationary vector autoregressive process:

$$\begin{aligned} \mathbf{h}_{t+1} = \mu + \mathbf{\Phi}(\mathbf{h}_t - \mu) + \eta_t, \quad \eta_t | \mathbf{h}_t \sim \mathcal{N}_p(0, \mathbf{\Sigma}_{\eta\eta}),\ t = 1, \ldots, n-1, \\ \mathbf{h}_1 \sim \mathcal{N}_p\left(\mu, \mathbf{\Sigma}_0\right), \end{aligned}$$

where $\mathbf{h}_t = (h_{1t}, \ldots, h_{pt})'$. To reduce the computational load, especially when p is large, the log volatilities can be assumed to be conditionally independent. In that case,

$$\begin{aligned} \mathbf{\Phi} &= \mathrm{diag}(\phi_{11}, ..., \phi_{pp}) \text{ and} \\ \mathbf{\Sigma}_{\eta\eta} &= \mathrm{diag}(\sigma_{1,\eta\eta}, ..., \sigma_{p,\eta\eta}) \end{aligned}$$

are both diagonal matrices. We refer to the former specification as the *VAR(1)* model and the latter as the *IAR(1)* (for independent autoregressive) model. Beyond these differences, the various models primarily differ in the way in which the outcomes y_t are modeled. In one formulation, the outcomes are assumed to be generated as

$$\mathbf{y}_t = \mathbf{V}_t^{1/2} \varepsilon_t, \quad \mathbf{V}_t^{1/2} = \mathrm{diag}\left(\exp(h_{1t}/2), \ldots, \exp(h_{pt}/2)\right), \quad t = 1, \ldots, n,$$

with the additional assumptions that

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} | \mathbf{h}_t \sim \mathcal{N}_{2p}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\varepsilon\varepsilon} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma}_{\eta\eta} \end{pmatrix}$$

and $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ is a matrix in correlation (with units on the main diagonal). Thus, conditioned on $\mathbf{h}_t$, $\mathrm{Var}(\mathbf{y}_t) = \mathbf{V}_t^{1/2} \mathbf{\Sigma}_{\varepsilon\varepsilon} \mathbf{V}_t^{1/2}$ is time-varying (as required), but

the conditional correlation matrix is $\mathbf{\Sigma}_{\varepsilon\varepsilon}$, which is not time-varying. In the sequel we refer to this model as the *basic MSV* model.

A second approach for modeling the outcome process is via a latent factor approach. In this case, the outcome model is specified as

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{V}_t^{1/2}\varepsilon_t,\ \mathbf{V}_t^{1/2} = \text{diag}\left(\exp(h_{1t}/2), \ldots, \exp(h_{pt}/2)\right),$$

where $\mathbf{B}$ is a $p \times q$ matrix ($q \leq p$) called the loading matrix, and $\mathbf{f}_t = (f_{1t}, ..., f_{qt})$ is a $q \times 1$ latent factor at time t. For identification reasons, the loading matrix is subject to some restrictions (that we present later in the paper), and $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ is the identity matrix. The model is closed by assuming that the latent variables are distributed independently across time as

$$\mathbf{f}_t|\mathbf{h}_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}_t),$$

where

$$\mathbf{D}_t = \text{diag}\left(\exp(h_{p+1,t}), \ldots, \exp(h_{p+q,t})\right)$$

is a diagonal matrix that depends on additional latent variables $h_{p+k,t}$. The full set of log volatilities, namely,

$$\mathbf{h}_t = (h_{1t}, \ldots, h_{pt}, h_{p+1,t}, ..., h_{p+q,t}),$$

are assumed to follow a VAR(1) or IAR(1) process. In this model, the variance of $\mathbf{y}_t$ conditional on the parameters and $\mathbf{h}_t$ is

$$\text{Var}(\mathbf{y}_t|\mathbf{h}_t) = \mathbf{V}_t + \mathbf{B}\mathbf{D}_t\mathbf{B}'$$

and as a result the conditional correlation matrix is time-varying.

Another way to model time-varying correlations is by direct modeling of the variance matrix $\mathbf{\Sigma}_t = \text{Var}(\mathbf{y}_t)$. One such model is the Wishart process model proposed by Philipov and Glickman (2006b), who assume that

$$\begin{aligned}\mathbf{y}_t|\mathbf{\Sigma}_t &\sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_t),\\ \mathbf{\Sigma}_t|\nu, \mathbf{S}_{t-1} &\sim \mathcal{IW}_p(\nu, \mathbf{S}_{t-1}),\end{aligned}$$

where $\mathcal{IW}_p(\nu_0, \mathbf{Q}_0)$ denotes a p-dimensional inverted Wishart distribution with parameters $(\nu_0, \mathbf{Q}_0)$, and $\mathbf{S}_{t-1}$ is a function of $\mathbf{\Sigma}_{t-1}$. Several models along these lines have been proposed as we discuss in Section 4.

The rest of the article is organized as follows. In Section 2, we first discuss the basic MSV model along with some of its extensions. Section 3 is devoted to the class of factor MSV models, while Section 4 deals with models in which the dynamics of the covariance matrix are modeled directly and Section 5 has our conclusions.

2 Basic MSV Model

2.1 No-leverage model

As in the preceding section, let $\mathbf{y}_t = (y_{1t}, \ldots, y_{pt})'$ denote a set of observations at time t on p financial variables and let $\mathbf{h}_t = (h_{1t}, \ldots, h_{pt})'$ be the corresponding vector of log volatilities. Then one approach to modeling the conditional covariance matrix of $\mathbf{y}_t$ is to assume that

$$\mathbf{y}_t = \mathbf{V}_t^{1/2}\varepsilon_t, \quad t = 1, \ldots, n, \tag{5}$$

$$\mathbf{h}_{t+1} = \mu + \mathbf{\Phi}(\mathbf{h}_t - \mu) + \eta_t, \quad t = 1, \ldots, n-1, \tag{6}$$

$$\mathbf{h}_1 \sim \mathcal{N}_p\left(\mu, \mathbf{\Sigma}_0\right), \tag{7}$$

where

$$\mathbf{V}_t^{1/2} = \operatorname{diag}\left(\exp(h_{1t}/2), \ldots, \exp(h_{pt}/2)\right),$$

$$\mu = (\mu_1, \ldots, \mu_p)'$$

and

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} | \mathbf{h}_t \sim \mathcal{N}_{2p}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\varepsilon\varepsilon} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\eta\eta} \end{pmatrix}.$$

Of course, for identification purposes, the diagonal elements of $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ must be 1, which means that the matrix $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ is a correlation matrix.

Analyses of this model are given by Harvey et al. (1994), Daníelsson (1998), Smith and Pitts (2006) and Chan et al. (2006). Actually, Harvey et al. (1994) dealt with a special case of this model in which $\mathbf{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_p)$. To fit the model, the measurement equation (5) is linearized by letting $w_{it} = \log y_{it}^2$. Because

$$\mathrm{E}(\log \varepsilon_{it}^2) = -1.27, \quad \mathrm{Var}(\log \varepsilon_{it}^2) = \pi^2/2, \tag{8}$$

one now has (a non-Gaussian) linear measurement equation:

$$\mathbf{w}_t = (-1.27)\mathbf{1} + \mathbf{h}_t + \xi_t, \tag{9}$$

where $\mathbf{w}_t = (w_{1t}, \ldots, w_{pt})'$, $\xi_t = (\xi_{1t}, \ldots, \xi_{pt})'$, $\xi_{it} = \log \varepsilon_{it}^2 + 1.27$ and $\mathbf{1} = (1, \ldots, 1)'$. Although the new state error ξ_t does not follow a normal distribution, approximate or quasi maximum likelihood (QML) estimates can be obtained by assuming Gaussianity. Calculation of the (misspecified) Gaussian likelihood also requires the covariance matrix of ξ_t. Harvey et al. (1994) showed that the (i, j)th element of the covariance matrix of $\xi_t = (\xi_{1t}, \ldots, \xi_{pt})'$ is given by $(\pi^2/2)\rho_{ij}^*$, where $\rho_{ii}^* = 1$ and

$$\rho_{ij}^* = \frac{2}{\pi^2} \sum_{n=1}^{\infty} \frac{(n-1)!}{\{\prod_{k=1}^{n}(1/2 + k - 1)\}n} \rho_{ij}^{2n}. \tag{10}$$

The model was applied to four daily foreign exchange rates (pound/dollar, Deutschemark/dollar, yen/dollar and Swiss franc/dollar). As mentioned in Harvey et al. (1994), the preceding fitting method cannot be extended to the leverage model considered below.

So et al. (1997) provide a similar analysis, but unlike Harvey et al. (1994) the nondiagonal elements of $\mathbf{\Phi}$ are not assumed to equal zero. Estimation of the parameters is again by the QML method which is implemented through a computationally efficient and numerically well-behaved expectation-maximization (EM) algorithm. The asymptotic variance-covariance matrix of the resulting estimates is based on the information matrix. Another related contribution is that of Daníelsson (1998), where the model

$$\begin{aligned}\mathbf{y}_t &= \mathbf{V}_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon}),\\ \mathbf{h}_{t+1} &= \mu + \operatorname{diag}(\phi_1, \ldots, \phi_p)(\mathbf{h}_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta})\end{aligned}$$

is analyzed. The parameters of this model are estimated by the simulated maximum likelihood method. The model and fitting method should be applied in the estimation of a bivariate model for foreign exchange rates (Deutschemark/dollar, yen/dollar) and stock indices (S&P500 and Tokyo stock exchange). On the basis of the log-likelihood values, they concluded that the MSV model is superior to alternative GARCH models such as the vector GARCH, diagonal vector GARCH (Bollerslev et al. (1988)), Baba–Engle–Kraft–Kroner (BEKK) model (Engle and Kroner (1995)) and the constant conditional correlation (CCC) model (Bollerslev (1990)).

Smith and Pitts (2006) considered a bivariate model without leverage that is similar to the model of Daníelsson (1998). The model is given by

$$\begin{aligned}\mathbf{y}_t &= \mathbf{V}_t^{1/2}\varepsilon_t, \quad \mathbf{V}_t^{1/2} = \operatorname{diag}(\exp(h_{1t}/2), \exp(h_{2t}/2)), \quad \varepsilon_t \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon}),\\ \mathbf{h}_{t+1} &= \mathbf{Z}_t\alpha + \operatorname{diag}(\phi_1, \phi_2)(\mathbf{h}_t - \mathbf{Z}_{t-1}\alpha) + \eta_t, \quad \eta_t \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}),\\ \mathbf{h}_1 &\sim \mathcal{N}_2(\mathbf{Z}_1\alpha_1, \mathbf{\Sigma}_0),\end{aligned}$$

where the (i,j)th element of $\mathbf{\Sigma}_0$ is the (i,j)th element of $\mathbf{\Sigma}_{\eta\eta}$ divided by $1 - \phi_i\phi_j$ to enforce the stationarity of $\mathbf{h}_t - \mathbf{Z}_t\alpha$. To measure the effect on daily returns in the yen/dollar foreign exchange of intervention by the Bank of Japan, they included in $\mathbf{Z}_t$ a variable that represents central bank intervention which they modeled by a threshold model. The resulting model was fit by Bayesian MCMC methods (Chib and Greenberg (1996), Chib (2001)). Because the likelihood of the parameters is complex, sampling of the posterior distributions in all applications of MCMC methods in MSV models is indirectly achieved by sampling the posterior distribution of the parameters and each of the latent variables. This tactic circumvents the computation of the likelihood. For this tactic to work it is necessary to efficiently sample the resulting high-dimensional posterior distribution. This is the challenge that has to be surmounted on a model-by-model basis.

To improve the efficiency of the MCMC algorithm, Smith and Pitts (2006) sampled $\mathbf{h}_t$'s in blocks, as in Shephard and Pitt (1997); see also Watanabe and Omori (2004). For simplicity, we describe their algorithm without the threshold specification and without missing observations. Let $Y_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ denote the set of observations until time t. Then the MCMC algorithm of Smith and Pitts (2006) is given by:

1. Sample $\{\mathbf{h}_t\}_{t=1}^n | \rho_{12}, \phi_1, \phi_2, \alpha, \boldsymbol{\Sigma}_{\eta\eta}, Y_n$. Divide $\{\mathbf{h}_t\}_{t=1}^n$ into several blocks, and sample a block at a time given other blocks. Let $\mathbf{h}_{a:b} = (\mathbf{h}_a', \ldots, \mathbf{h}_b')'$. To sample a block $\mathbf{h}_{a:b}$ given other $\mathbf{h}_j$'s, we conduct a Metropolis–Hastings (M-H) algorithm using a proposal density of the type introduced by Chib and Greenberg (1994, 1998) and Chib (2001):

$$\mathbf{h}_{a:b} \sim \mathcal{N}_{2(b-a+1)} \left(\hat{\mathbf{h}}_{a:b}, \left[-\frac{\partial l(\mathbf{h}_{a:b})}{\partial \mathbf{h}_{a:b} \partial \mathbf{h}_{a:b}'} \right]^{-1}_{\mathbf{h}_{a:b} = \hat{\mathbf{h}}_{a:b}} \right),$$

   where

$$l(\mathbf{h}_{a:b}) = \text{const} - \frac{1}{2} \cdot \sum_{t=a}^{b} \left(\mathbf{1}' \mathbf{h}_t + \mathbf{y}_t' \mathbf{V}_t^{-1/2} \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{V}_t^{-1/2} \mathbf{y}_t \right) - \frac{1}{2} \cdot$$
$$\sum_{t=a}^{b+1} \{\mathbf{h}_t - \mathbf{Z}_t \alpha - \boldsymbol{\Phi}(\mathbf{h}_{t-1} - \mathbf{Z}_{t-1}\alpha)\}' \boldsymbol{\Sigma}_{\eta\eta}^{-1} \{\mathbf{h}_t - \mathbf{Z}_t \alpha - \boldsymbol{\Phi}(\mathbf{h}_{t-1} - \mathbf{Z}_{t-1}\alpha)\}.$$

   The proposal density is a Gaussian approximation of the conditional posterior density based on a Taylor expansion of the conditional posterior density around the mode $\hat{\mathbf{h}}_{a:b}$. The mode is found numerically by the Newton–Raphson method.
2. Sample $\rho_{12} | \{\mathbf{h}_t\}_{t=1}^n, \phi_1, \phi_2, \alpha, \boldsymbol{\Sigma}_{\eta\eta}, Y_n$ using the M-H algorithm.
3. Sample $\phi_1, \phi_2 | \{\mathbf{h}_t\}_{t=1}^n, \rho_{12}, \alpha, \boldsymbol{\Sigma}_{\eta\eta}, Y_n$ using the M-H algorithm.
4. Sample $\alpha | \{\mathbf{h}_t\}_{t=1}^n, \rho_{12}, \phi_1, \phi_2, \boldsymbol{\Sigma}_{\eta\eta}, Y_n \sim \mathcal{N}_2(\delta, \boldsymbol{\Sigma})$, where

$$\delta = \boldsymbol{\Sigma} \sum_{t=2}^{n} (\mathbf{Z}_t - \boldsymbol{\Phi}\mathbf{Z}_{t-1})' \boldsymbol{\Sigma}_{\eta\eta}^{-1} (\mathbf{h}_t - \boldsymbol{\Phi}\mathbf{h}_{t-1}) + \mathbf{Z}_1' \boldsymbol{\Sigma}_0^{-1} \mathbf{h}_1,$$
$$\boldsymbol{\Sigma}^{-1} = \sum_{t=2}^{n} (\mathbf{Z}_t - \boldsymbol{\Phi}\mathbf{Z}_{t-1})' \boldsymbol{\Sigma}_{\eta\eta}^{-1} (\mathbf{Z}_t - \boldsymbol{\Phi}\mathbf{Z}_{t-1}) + \mathbf{Z}_1' \boldsymbol{\Sigma}_0^{-1} \mathbf{Z}_1.$$

5. Sample $\boldsymbol{\Sigma}_{\eta\eta} | \{\mathbf{h}_t\}_{t=1}^n, \rho_{12}, \phi_1, \phi_2, \alpha, Y_n$ using the M-H algorithm.

Bos and Shephard (2006) considered a similar model but with the mean in the outcome specification driven by an $r \times 1$ latent process vector α_t:

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{Z}_t\alpha_t + \mathbf{G}_t\mathbf{u}_t,\\
\alpha_{t+1} &= \mathbf{T}_t\alpha_t + \mathbf{H}_t\mathbf{u}_t,\\
\mathbf{u}_t &= \mathbf{V}_t^{1/2}\varepsilon_t, \quad \mathbf{V}_t^{1/2} = \mathrm{diag}(\exp(h_{1t}/2),\ldots,\exp(h_{qt}/2)), \quad \varepsilon_t \sim \mathcal{N}_q(\mathbf{0},\mathbf{I}),\\
\mathbf{h}_{t+1} &= \mu + \mathbf{\Phi}(\mathbf{h}_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}_q(\mathbf{0},\mathbf{\Sigma}_{\eta\eta}), \quad \mathbf{h}_t = (h_{1t},\ldots,h_{qt})',
\end{aligned}
$$

where $\mathbf{G}_t\mathbf{u}_t$ and $\mathbf{H}_t\mathbf{u}_t$ are independent and the off-diagonal element of $\mathbf{\Phi}$ may be nonzero. Given $\{\mathbf{h}_t\}_{t=1}^n$, this is a linear Gaussian state-space model,

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{Z}_t\alpha_t + \mathbf{u}_t^*, \quad \mathbf{u}_t^* \sim \mathcal{N}_p(\mathbf{0}, \mathbf{G}_t\mathbf{V}_t\mathbf{G}_t'),\\
\alpha_{t+1} &= \mathbf{T}_t\alpha_t + \mathbf{v}_t^*, \quad \mathbf{v}_t^* \sim \mathcal{N}_r(\mathbf{0}, \mathbf{H}_t\mathbf{V}_t\mathbf{H}_t'),
\end{aligned}
$$

where $\mathbf{u}_t^*$ and $\mathbf{v}_t^*$ are independent. Bos and Shephard (2006) took a Bayesian approach and conducted the MCMC simulation in two blocks. Let $\theta = (\psi, \lambda)$, where ψ indexes the unknown parameters in $\mathbf{T}_t, \mathbf{Z}_t, \mathbf{G}_t, \mathbf{H}_t$, and λ denotes the parameter of the SV process of $\mathbf{u}_t$.

1. Sample $\theta, \{\alpha_t\}_{t=1}^n | \{\mathbf{h}_t\}_{t=1}^n, Y_n$.
   - Sample $\theta | \{\mathbf{h}_t\}_{t=1}^n, Y_n$ using a M-H algorithm or a step from the adaptive rejection Metropolis sampler by Gilks et al. (1995); see Bos and Shephard (2006).
   - Sample $\{\alpha_t\}_{t=1}^n | \theta, \{\mathbf{h}_t\}_{t=1}^n, Y_n$ using a simulation smoother for a linear Gaussian state-space model; see de Jong and Shephard (1995) and Durbin and Koopman (2002). We first sample disturbances of the linear Gaussian state-space model and obtain samples of α_t recursively.

2. Sample $\{\mathbf{h}_t\}_{t=1}^n | \theta, \{\alpha_t\}_{t=1}^n, Y_n$. For $t = 1,\ldots,n$, we sample $\mathbf{h}_t$ one at a time by the M-H algorithm with the proposal distribution

$$
\begin{aligned}
\mathbf{h}_t | \mathbf{h}_{t-1}, \mathbf{h}_{t+1}, \theta &\sim \mathcal{N}_q\left(\mu + \mathbf{Q}\mathbf{\Phi}'\mathbf{\Sigma}_{\eta\eta}^{-1}\left\{(\mathbf{h}_{t+1} - \mu) + (\mathbf{h}_{t-1} - \mu)\right\}, \mathbf{Q}\right),\\
&\qquad\qquad t = 2,\ldots,n-1,\\
\mathbf{h}_n | \mathbf{h}_{n-1}, \theta &\sim \mathcal{N}_q(\mu, \mathbf{\Sigma}_{\eta\eta}),
\end{aligned}
$$

   where $\mathbf{Q}^{-1} = \mathbf{\Sigma}_{\eta\eta}^{-1} + \mathbf{\Phi}'^{-1}\mathbf{\Phi}$.

Although the sampling scheme which samples $\mathbf{h}_t$ at a time is expected to produce highly autocorrelated MCMC samples, the adaptive rejection Metropolis sampling of θ seems to overcome some of the inefficiencies. Yu and Meyer (2006) provide a survey of MSV models that proceed along these lines and illustrate how the Bayesian software program WinBUGS can be used to fit bivariate models.

It is worth mentioning that it is possible to relax the assumption that the volatility process is VAR of order 1. In one notable attempt, So and Kwok (2006) consider a MSV model where the volatility vector $\mathbf{h}_t - \mu$ follows a stationary vector autoregressive fractionally integrated moving average process,

ARFIMA$(p, \mathbf{d}, q)$, such that

$$\mathbf{\Phi}(B)D(B)(\mathbf{h}_{t+1} - \mu) = \Theta(B)\eta_t, \quad \eta_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}), \tag{11}$$

$$D(B) = \mathrm{diag}((1-B)^{d_1}, \ldots, (1-B)^{d_p}), \quad |d_i| < 1/2, \tag{12}$$

$$\mathbf{\Phi}(B) = \mathbf{I} - \mathbf{\Phi}_1 B - \cdots - \mathbf{\Phi}_p B^p, \tag{13}$$

$$\Theta(B) = \mathbf{I} + \mathbf{\Theta}_1 B + \cdots + \mathbf{\Theta}_q B^q, \tag{14}$$

where B is a backward operator such that $B^j \mathbf{h}_t = \mathbf{h}_{t-j}$. The ε_t and η_t are assumed to be independent. So and Kwok (2006) investigated statistical properties of the model and proposed a QML estimation method as in Harvey et al. (1994). They linearized the measurement equation by taking the logarithm of the squared returns and considered the linear state-space model

$$\begin{aligned} \mathbf{w}_t &= (-1.27)\mathbf{1} + \mathbf{h}_t + \xi_t, \\ \mathbf{\Phi}(B)D(B)(\mathbf{h}_{t+1} - \mu) &= \Theta(B)\eta_t, \end{aligned}$$

where $\mathbf{w}_t = (w_{1t}, \ldots, w_{pt})'$, $\xi_t = (\xi_{1t}, \ldots, \xi_{pt})'$, $w_{it} = \log y_{it}^2$, and $\xi_{it} = \log \varepsilon_{it}^2$ for $i = 1, \ldots, n$. The covariance matrix of ξ_t can be obtained as in Harvey et al. (1994). To conduct the QML estimation, So and Kwok (2006) assumed that ξ_t follows a normal distribution and obtained estimates based on the linear Gaussian state-space model. However, since $\mathbf{h}_t - \mu$ follows a vector ARFIMA$(p, \mathbf{d}, q)$ process, the conventional Kalman filter is not applicable as the determinant and inverse of a large covariance matrix is required to calculate the quasi-log-likelihood function. To avoid this calculation, So and Kwok (2006) approximated the quasi-log-likelihood function by using a spectral likelihood function based on a Fourier transform.

2.2 Leverage effects

Another extension of the basic MSV model is to allow for correlation between ε_t and η_t by letting $\mathbf{\Sigma}_{\varepsilon\eta} \neq \mathbf{O}$. This extension is important because at least for returns on stocks there is considerable evidence that the measurement and volatility innovations are correlated (Yu (2005), Omori et al. (2007)). That this correlation (the leverage effect) should be modeled is mentioned by Daníelsson (1998) but this suggestion is not implemented in his empirical study of foreign exchange rates and stock indices. One compelling work on a type of leverage model is due to Chan et al. (2006), who considered the model

$$\begin{aligned} \mathbf{y}_t &= \mathbf{V}_t^{1/2} \varepsilon_t, \\ \mathbf{h}_{t+1} &= \mu + \mathrm{diag}(\phi_1, \ldots, \phi_p)(\mathbf{h}_t - \mu) + \mathbf{\Psi}^{1/2} \eta_t, \\ \mathbf{h}_1 &\sim \mathcal{N}_p(\mu, \mathbf{\Psi}^{1/2} \mathbf{\Sigma}_0 \mathbf{\Psi}^{1/2}), \end{aligned}$$

where the (i,j) element of $\mathbf{\Sigma}_0$ is the (i,j) element of $\mathbf{\Sigma}_{\eta\eta}$ divided by $1-\phi_i\phi_j$ satisfying a stationarity condition such that

$$\mathbf{\Sigma}_0 = \mathbf{\Phi}\mathbf{\Sigma}_0\mathbf{\Phi} + \mathbf{\Sigma}_{\eta\eta}$$

and

$$\begin{aligned}
\mathbf{V}_t^{1/2} &= \operatorname{diag}\left(\exp(h_{1t}/2), \ldots, \exp(h_{pt}/2)\right), \\
\mathbf{\Psi}^{1/2} &= \operatorname{diag}\left(\sqrt{\psi_1^2}, \ldots, \sqrt{\psi_p^2}\right), \\
\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} &\sim \mathcal{N}_{2p}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\varepsilon\varepsilon} & \mathbf{\Sigma}_{\varepsilon\eta} \\ \mathbf{\Sigma}_{\eta\varepsilon} & \mathbf{\Sigma}_{\eta\eta} \end{pmatrix}.
\end{aligned}$$

Actually, the model considered in Chan et al. (2006) had correlation between ε_t and η_{t-1}, which is not correctly a model of leverage. Our discussion therefore modifies their treatment to deal with the model just presented, where ε_t and η_t are correlated. Note that $\mathbf{\Sigma}$ is a $2p \times 2p$ correlation matrix with $\mathbf{\Sigma}_{\varepsilon\eta} \neq \mathbf{O}$. Now, following Wong et al. (2003) and Pitt et al. (2006), we reparameterize $\mathbf{\Sigma}$ such that

$$\mathbf{\Sigma}^{-1} = \mathbf{TGT}, \quad \mathbf{T} = \operatorname{diag}\left(\sqrt{G^{11}}, \ldots, \sqrt{G^{pp}}\right),$$

where $\mathbf{G}$ is a correlation matrix and G^{ii} denotes the (i,i)th element of the inverse matrix of $\mathbf{G}$. Under this parameterization, we can find the posterior probability that the strict lower triangle of the transformed correlation matrix $\mathbf{G}$ is equal to zero. Let $J_{ij} = 1$ if $G_{ij} \neq 0$ and $J_{ij} = 0$ if $G_{ij} = 0$ for $i = 1, \ldots, 2p, j < i$ and $S(\mathbf{J})$ denote the number of elements that are 1's in $\mathbf{J} = \{J_{ij}, i = 1, \ldots, 2p,\ j < i\}$. Further let $\mathbf{G}_{\{J=k\}} = \{G_{ij} : J_{ij} = k \in \mathbf{J}\}$ $(k = 0, 1)$ and $\mathcal{A}$ denote a class of $2p \times 2p$ correlation matrices. Wong et al. (2003) proposed a hierarchical prior for $\mathbf{G}$:

$$\begin{aligned}
\pi(d\mathbf{G}|\mathbf{J}) &= V(\mathbf{J})^{-1} d\mathbf{G}_{\{J=1\}} I(\mathbf{G} \in \mathcal{A}), \quad V(\mathbf{J}) = \int_{\mathbf{G}\in\mathcal{A}} d\mathbf{G}_{\{J=1\}}, \\
\pi(\mathbf{J}|S(\mathbf{J}) = l) &= \frac{V(\mathbf{J})}{\displaystyle\sum_{\mathbf{J}^*: S(\mathbf{J}^*)=l} V(\mathbf{J}^*)}, \\
\pi(S(\mathbf{J}) = l|\varphi) &= \binom{p(2p-1)}{l} \varphi^l (1-\varphi)^{p(2p-1)-l}.
\end{aligned}$$

If we assume $\varphi \sim \mathcal{U}(0,1)$, the marginal prior probability $\pi(S(\mathbf{J}) = l) = 1/(p(2p-1)+1)$; see Wong et al. (2003) for the evaluation of $V(\mathbf{J})$. Let $\phi = (\phi_1, \ldots, \phi_p)'$ and $\psi = (\psi_1, \ldots, \psi_p)'$ $(\psi_j > 0, j = 1, \ldots, p)$.

1. Sample $\phi|\mu, \{\mathbf{h}_t\}_{t=1}^n, \psi, \mathbf{\Sigma}, Y_n$ where $Y_n = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Let $\mathbf{\Sigma}^{ij}$ denote the (i,j)th block of the $2p \times 2p$ matrix $\mathbf{\Sigma}^{-1}$ and $\mathbf{d}$ be a vector that consists

of the diagonal elements

$$\sum_{t=1}^{n-1} \boldsymbol{\Psi}^{-1/2}(\mathbf{h}_t - \mu)\left(\mathbf{y}_t'\mathbf{V}_t^{-1/2}\boldsymbol{\Sigma}^{12} + \boldsymbol{\Psi}^{-1/2}(\mathbf{h}_{t+1} - \mu)'\boldsymbol{\Sigma}^{22}\right).$$

Propose a candidate

$$\begin{aligned}
\phi &\sim \mathcal{TN}_R(\mu_\phi, \boldsymbol{\Sigma}_\phi), \quad R = \{\phi : \phi_j \in (-1,1), j = 1,\ldots,p\},\\
\boldsymbol{\Sigma}_\phi^{-1} &= \boldsymbol{\Sigma}^{22} \odot \left\{\sum_{t=1}^{n-1} \boldsymbol{\Psi}^{-1/2}(\mathbf{h}_t - \mu)(\mathbf{h}_t - \mu)'\boldsymbol{\Psi}^{-1/2}\right\},\\
\mu_\phi &= \boldsymbol{\Sigma}_\phi \mathbf{d},
\end{aligned}$$

where $\odot$ is the element-by-element multiplication operator (Hadamard product) and apply the M-H algorithm.

2. Sample $\mu|\phi, \{\mathbf{h}_t\}_{t=1}^n, \psi, \boldsymbol{\Sigma}, Y_n \sim \mathcal{N}_p(\mu_*, \boldsymbol{\Sigma}_*)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_*^{-1} &= (n-1)(\mathbf{I} - \boldsymbol{\Phi})\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}^{22}\boldsymbol{\Psi}^{-1/2}(\mathbf{I} - \boldsymbol{\Phi}) + \boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Psi}^{-1/2},\\
\mu_* &= \boldsymbol{\Sigma}_*\left[(\mathbf{I} - \boldsymbol{\Phi})\boldsymbol{\Psi}^{-1/2}\sum_{t=1}^{n-1}\left\{\boldsymbol{\Sigma}^{21}\mathbf{V}_t^{-1/2}\mathbf{y}_t + \boldsymbol{\Sigma}^{22}\boldsymbol{\Psi}^{-1/2}(\mathbf{h}_{t+1} - \boldsymbol{\Phi}\mathbf{h}_t)\right\}\right.\\
&\quad \left. + \boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Psi}^{-1/2}\mathbf{h}_1\right].
\end{aligned}$$

3. Sample $\psi|\phi, \mu, \{\mathbf{h}_t\}_{t=1}^n, \boldsymbol{\Sigma}, Y_n$. Let $\mathbf{v} = (\psi_1^{-1}, \ldots, \psi_p^{-1})$ and $l(\mathbf{v})$ denote the logarithm of the conditional probability density of $\mathbf{v}$ and $\hat{\mathbf{v}}$ denote the mode of $l(\mathbf{v})$. Then conduct the M-H algorithm using a truncated multivariate t distribution on the region $R = \{\mathbf{v} : v_j > 0, j = 1,\ldots,p\}$ with six degrees of freedom, location parameter $\hat{\mathbf{v}}$ and a covariance matrix $-\{\partial^2 l(\mathbf{v})/\partial\mathbf{v}\partial\mathbf{v}'\}_{\mathbf{v}=\hat{\mathbf{v}}}^{-1}$.
4. Sample $\{\mathbf{h}_t\}_{t=1}^n|\phi, \mu, \psi, \boldsymbol{\Sigma}, Y_n$. We divide $\{\mathbf{h}_t\}_{t=1}^n$ into several blocks, and sample a block at a time given other blocks as in Smith and Pitts (2006). Let $\mathbf{h}_{a:b} = (\mathbf{h}_a', \ldots, \mathbf{h}_b')'$. To sample a block $\mathbf{h}_{a:b}$ given other $\mathbf{h}_j$'s, we conduct a M-H algorithm using a Chib and Greenberg (1994) proposal,

$$\begin{aligned}
\mathbf{h}_{a:b} &\sim \mathcal{N}_{p(b-a+1)}\left(\hat{\mathbf{h}}_{a:b}, \left[-\frac{\partial l(\mathbf{h}_{a:b})}{\partial\mathbf{h}_{a:b}\partial\mathbf{h}_{a:b}'}\right]_{\mathbf{h}_{a:b}=\hat{\mathbf{h}}_{a:b}}^{-1}\right),\\
l(\mathbf{h}_{a:b}) &= \text{const} - \frac{1}{2}\sum_{t=a}^{b}\mathbf{1}'\mathbf{h}_t - \frac{1}{2}\sum_{t=a}^{b+1}\mathbf{r}_t'\boldsymbol{\Sigma}^{-1}\mathbf{r}_t,\\
\mathbf{r}_t &= \begin{pmatrix}\mathbf{V}_t^{-1/2}\mathbf{y}_t\\ \boldsymbol{\Psi}^{-1/2}\{\mathbf{h}_{t+1} - \mu - \boldsymbol{\Phi}(\mathbf{h}_t - \mu)\}\end{pmatrix},
\end{aligned}$$

a Gaussian approximation of the conditional posterior density based on Taylor expansion of the conditional posterior density around the mode $\hat{\mathbf{h}}_{a:b}$. The mode is found using the Newton–Raphson method numerically. The analytical derivatives can be derived similarly as in the Appendix of Chan et al. (2006).

5. Sample $\boldsymbol{\Sigma}|\phi, \mu, \psi, \{\mathbf{h}_t\}_{t=1}^n, Y_n$. Using the parsimonious reparameterization proposed in Wong et al. (2003), we generate each element G_{ij} one at a time using the M-H algorithm.

Chan et al. (2006) applied the proposed estimation method to equities at three levels of aggregation: (1) returns for eight different markets (portfolios of stocks in NYSE, AMEX, NASDAQ and S&P500 indices); (2) returns for eight different industries (portfolios of eight well-known and actively traded stocks in petroleum, food products, pharmaceutical, banks, industrial equipment, aerospace, electric utilities, and department/discount stores); (3) returns for individual firms within the same industry. They found strong evidence of correlation between ε_t and η_{t-1} only for the returns of the eight different markets and suggested that this correlation is mainly a feature of marketwide rather than firm-specific returns and volatility.

Asai and McAleer (2006) also analyzed a MSV model with leverage effects, letting

$$\begin{aligned}\boldsymbol{\Phi} &= \operatorname{diag}(\phi_1, \ldots, \phi_p),\\ \boldsymbol{\Sigma}_{\varepsilon\eta} &= \operatorname{diag}\left(\lambda_1\sigma_{1,\eta\eta}, \ldots, \lambda_p\sigma_{p,\eta\eta}\right).\end{aligned}$$

The cross-asset leverage effects are assumed to be 0 ($\operatorname{Corr}(\varepsilon_{it}, \eta_{jt}) = 0$, for $i \neq j$). As in Harvey and Shephard (1996), they linearized the measurement equations and considered the following state-space model conditional on $\mathbf{s}_t = (s_{1t}, \ldots, s_{pt})'$, where $s_{it} = 1$ if y_{it} is positive and $s_{it} = -1$ otherwise:

$$\begin{aligned}\log y_{it}^2 &= h_{it} + \zeta_{it}, \quad \zeta_{it} = \log \varepsilon_{it}^2, \quad i = 1, \ldots, p, \quad t = 1, \ldots, n,\\ \mathbf{h}_{t+1} &= \tilde{\mu} + \mu_t^* + \operatorname{diag}(\phi_1, \ldots, \phi_p)\mathbf{h}_t + \eta_t^*,\\ \mu_t^* &= \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}_{\varepsilon\eta}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}\mathbf{s}_t, \quad \eta_t^* \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\eta_t^*\eta_t^*}),\end{aligned}$$

where $\mathrm{E}(\zeta_{it}) = -1.27$, and $\operatorname{Cov}(\zeta_{it}, \zeta_{jt}) = (\pi^2/2)\rho_{ij}^*$ given in (10). The matrix $\boldsymbol{\Sigma}_{\eta_t^*\eta_t^*}$ and $\mathrm{E}(\eta_t^*\zeta_t')$ are given in Asai and McAleer (2006). They also considered an alternative MSV model with leverage effects and size effects given by

$$\begin{aligned}\mathbf{h}_{t+1} &= \tilde{\mu} + \boldsymbol{\Gamma}_1\mathbf{y}_t + \boldsymbol{\Gamma}_2|\mathbf{y}_t| + \boldsymbol{\Phi}\mathbf{h}_t + \eta_t,\\ \boldsymbol{\Gamma}_1 &= \operatorname{diag}(\gamma_{11}, \ldots, \gamma_{1p}), \quad \boldsymbol{\Gamma}_2 = \operatorname{diag}(\gamma_{21}, \ldots, \gamma_{2p}),\\ |\mathbf{y}_t| &= (|y_{1t}|, \ldots, |y_{pt}|)', \qquad \boldsymbol{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_p),\\ \boldsymbol{\Sigma}_{\varepsilon\eta} &= \mathbf{O}.\end{aligned}$$

This model is a generalization of a univariate model given by Daníelsson (1994). It incorporates both leverage effects and the magnitude of the previous returns through their absolute values. Asai and McAleer (2006) fit these two models to returns of three stock indices—S&P500 Composite Index, the Nikkei 225 Index, and the Hang Seng Index—by an importance sampling Monte Carlo maximum likelihood estimation method. They found that the MSV model with leverage and size effects is preferred in terms of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) measures.

2.3 Heavy-tailed measurement error models

It has by now been quite well established that the tails of the distribution of asset returns are heavier than those of the Gaussian. To deal with this situation it has been popular to employ the Student t distribution as a replacement for the default Gaussian assumption. One reason for the popularity of the Student t distribution is that it has a simple hierarchical form as a scale mixture of normals. Specifically, if T is distributed as standard Student t with ν degrees of freedom then T can be expressed as

$$T = \lambda^{-1/2} Z, \quad Z \sim \mathcal{N}(0,1), \quad \lambda \sim \mathcal{G}(\nu/2, \nu/2).$$

This representation can be exploited in the fitting, especially in the Bayesian context. One early example of the use of the Student t distribution occurred in Harvey et al. (1994), who assumed that in connection with the measurement error ε_{it} that

$$\epsilon_{it} = \lambda_{it}^{-1/2} \varepsilon_{it}, \quad \varepsilon_t \sim \text{ i.d.d. } \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon}), \quad \lambda_{it} \sim \text{ i.d.d. } \mathcal{G}(\nu_i/2, \nu_i/2),$$

where the mean is $\mathbf{0}$ and the elements of the covariance matrix are given by

$$\operatorname{Cov}(\epsilon_{it}, \epsilon_{jt}) = \begin{cases} \dfrac{\nu_i}{\nu_i - 2}, & i = j, \\ \mathrm{E}(\lambda_{it}^{-1/2}) \mathrm{E}(\lambda_{jt}^{-1/2}) \rho_{ij}, & i \neq j, \end{cases}$$

$$\text{and} \quad \mathrm{E}(\lambda_{it}^{-1/2}) = \frac{(\nu_i/2)^{1/2} \Gamma((\nu_i - 1)/2)}{\Gamma(\nu_i/2)}.$$

Alternatively, the model can now be expressed as

$$\mathbf{y}_t = \mathbf{V}_t^{1/2} \mathbf{\Lambda}_t^{-1/2} \varepsilon_t, \quad \mathbf{\Lambda}_t^{-1/2} = \operatorname{diag}\left(1/\sqrt{\lambda_{1t}}, \ldots, 1/\sqrt{\lambda_{pt}}\right).$$

Taking the logarithm of squared ϵ_{it}, one gets

$$\log \epsilon_{it}^2 = \log \varepsilon_{it}^2 - \log \lambda_{it}.$$

They derived the QML estimators using the mean and covariance matrix of $(\log \epsilon_{it}^2, \log \epsilon_{jt}^2)$ using

$$\mathrm{E}(\log \lambda_{it}) = \psi'(\nu/2) - \log(\nu/2), \quad \mathrm{Var}(\log \lambda_{it}) = \psi'(\nu/2),$$

and (8) and (10), where ψ and ψ' are the digamma and trigamma functions. On the other hand, Yu and Meyer (2006) considered a multivariate Student t distribution for ε_t, in which case the measurement error has the form

$$\mathbf{T} = \lambda_t^{-1/2} \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}), \quad \lambda_t \sim \mathcal{G}(\nu/2, \nu/2).$$

They mentioned that this formulation was empirically better supported than the formulation in Harvey et al. (1994). The model was fit by Bayesian MCMC methods.

Another alternative to the Gaussian distribution is the generalized hyperbolic (GH) distribution introduced by Barndorff-Nielsen (1977). This family is also a member of the scale mixture of normals family of distributions. In this case, the mixing distribution is a generalized inverse Gaussian distribution. The GH distribution is a rich class of distributions that includes the normal, normal inverse Gaussian, reciprocal normal inverse Gaussian, hyperbolic, skewed Student's t, Laplace, normal gamma, and reciprocal normal hyperbolic distributions (Barndorff-Nielsen and Shephard (2001)). Aas and Haff (2006) employed the univariate GH distributions (normal inverse Gaussian distributions and univariate GH skew Student's t distributions) and estimated in the analysis of the total index of Norwegian stocks (TOTX), the SSBWG hedged bond index for international bonds, the Norwegian kroner/euro exchange rate and the EURIBOR five-year interest rate. They found that the GH skew Student's t distribution is superior to the normal inverse Gaussian distribution for heavy-tailed data, and superior to the skewed t distribution proposed by Azzalini and Capitanio (2003) for very skewed data.

The random variable $\mathbf{x} \sim \mathcal{GH}(\nu, \alpha, \beta, \mathbf{m}, \delta, \mathbf{S})$ follows a multivariate GH distribution with density

$$f(\mathbf{x}) = \frac{(\gamma/\delta)^\nu K_{\nu-\frac{p}{2}} \left(\alpha \sqrt{\delta^2 + (\mathbf{x}-\mathbf{m})' \mathbf{S}^{-1} (\mathbf{x}-\mathbf{m})} \right) \exp\{\beta'(\mathbf{x}-\mathbf{m})\}}{(2\pi)^{\frac{p}{2}} K_\nu(\delta\gamma) \left\{ \alpha^{-1} \sqrt{\delta^2 + (\mathbf{x}-\mathbf{m})' \mathbf{S}^{-1} (\mathbf{x}-\mathbf{m})} \right\}^{\frac{p}{2}-\nu}}, \tag{15}$$

$$\gamma \equiv \sqrt{\alpha^2 - \beta' \mathbf{S} \beta} \geq 0, \quad \alpha^2 \geq \beta' \mathbf{S} \beta,$$
$$\nu, \alpha \in R, \quad \beta, \mathbf{m} \in R^n, \quad \delta > 0,$$

where K_ν is a modified Bessel function of the third kind, and $\mathbf{S}$ is a $p \times p$ positive-definite matrix with determinant $|\mathbf{S}| = 1$ (Protassov (2004), Schmidt et al. (2006)). It can be shown that $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \mathbf{m} + z_t \mathbf{S} \beta + \sqrt{z}_t \mathbf{S}^{1/2} \varepsilon_t,$$

where $\mathbf{S}^{1/2}$ is a $p \times p$ matrix such that $\mathbf{S} = \mathbf{S}^{1/2}\mathbf{S}^{1/2'}$ and $\varepsilon \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and $z_t \sim \mathcal{GIG}(\nu, \delta, \gamma)$ follows a generalized inverse Gaussian distribution which we denote $z \sim \mathcal{GIG}(\nu, \delta, \gamma)$ and whose density is given by

$$f(z) = \frac{(\gamma/\delta)^\nu}{2K_\nu(\delta\gamma)} z^{\nu-1} \exp\left\{-\frac{1}{2}\left(\delta^2 z^{-1} + \gamma^2 z\right)\right\}, \quad \gamma, \delta \geq 0, \quad \nu \in R, \quad z > 0,$$

where the range of the parameters is given by

$$\begin{gathered}\delta > 0, \gamma^2 \geq 0, \text{ if } \nu < 0,\\ \delta > 0, \gamma^2 > 0, \text{ if } \nu = 0,\\ \delta \geq 0, \gamma^2 > 0, \text{ if } \nu > 0.\end{gathered}$$

For the generation of a random sample from $\mathcal{GIG}(\nu, a, b)$, see Dagpunar (1989), Doornik (2002) and Hörmann et al. (2004). The estimation of such a multivariate distribution would be difficult and Protassov (2004) relied on the EM algorithm with ν fixed and fit the five-dimensional normal inverse Gaussian distribution to a series of returns on foreign exchange rates (Swiss franc, Deutschemark, British pound, Canadian dollar and Japanese yen). Schmidt et al. (2006) proposed an alternative class of distributions, called the multivariate affine GH class, and applied it to bivariate models for various asset returns data (Dax, Cac, Nikkei and Dow returns). Other multivariate skew densities have also been proposed, for example, in Arellano-Valle and Azzalini (2006), Bauwens and Laurent (2005), Dey and Liu (2005) Azzalini (2005), Gupta et al. (2004) and Ferreira and Steel (2004).

3 Factor MSV Model

3.1 Volatility factor model

A weakness of the preceding MSV models is that the implied conditional correlation matrix does not vary with time. One approach for generating time-varying correlations is via factor models in which the factors follow a SV process. One type of factor SV model (that however does not lead to time-varying correlations) was considered by Quintana and West (1987) and by Jungbacker and Koopman (2006), who utilized a single factor to decompose the outcome into two multiplicative components, a scalar common volatility factor and a vector of idiosyncratic noise variables, as

$$\begin{aligned}\mathbf{y}_t &= \exp\left(\frac{h_t}{2}\right)\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon}),\\ h_{t+1} &= \mu + \phi(h_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2),\end{aligned}$$

where h_t is a scalar. The first element in $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ is assumed to be 1 for identification reasons. By construction, the positivity of the variance of $\mathbf{y}_t$ is ensured. In comparison with the basic MSV model, this model has fewer parameters, which makes it more convenient to fit. The downside of the model, however, is that unlike the mean factor MSV model which we discuss below, the conditional correlations in this model are time-invariant. Moreover, the correlation between log volatilities is 1, which is clearly limiting.

In order to estimate the model, Jungbacker and Koopman (2006) applied a Monte Carlo likelihood method to fit data on exchange rate returns of the British pound, the Deutschemark and the Japanese yen against the US dollar. They found that the estimate of ϕ is atypically low, indicating that the model is inappropriate for explaining the movements of multivariate volatility.

A more general version of this type was considered by Harvey et al. (1994), who introduced a common factor in the linearized state-space version of the basic MSV model by letting

$$\mathbf{w}_t = (-1.27)\mathbf{1} + \mathbf{\Theta}\mathbf{h}_t + \overline{\mathbf{h}} + \xi_t, \tag{16}$$

$$\mathbf{h}_{t+1} = \mathbf{h}_t + \eta_t, \quad \eta_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}), \tag{17}$$

where $\mathbf{w}_t = (w_{1t}, \ldots, w_{pt})'$, $\xi_t = (\xi_{1t}, \ldots, \xi_{pt})'$ and $\mathbf{h}_t = (h_{1t}, \ldots, h_{qt})'$ ($q \leq p$). Furthermore, one assumes that

$$\mathbf{\Theta} = \begin{pmatrix} \theta_{11} & 0 & \cdots & 0 \\ \theta_{21} & \theta_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \theta_{q1} & \cdots & \theta_{q,q-1} & \theta_{qq} \\ \vdots & & \vdots & \vdots \\ \theta_{p,1} & \cdots & \theta_{p,q-1} & \theta_{p,q} \end{pmatrix}, \quad \overline{\mathbf{h}} = \begin{pmatrix} \mathbf{0} \\ \overline{h}_{q+1} \\ \vdots \\ \overline{h}_p \end{pmatrix}.$$

Harvey et al. (1994) estimated the parameters by the QML method. To make the factor loadings interpretable, the common factors are rotated such that $\mathbf{\Theta}^* = \mathbf{\Theta}\mathbf{R}'$ and $\mathbf{h}_t^* = \mathbf{R}\mathbf{h}_t$, where $\mathbf{R}$ is an orthogonal matrix.

Tims and Mahieu (2006) considered a similar but simpler model for the logarithm of the range of the exchange rates in the context of an application involving four currencies. Let w_{ij} denote a logarithm of the range of foreign exchange rate of the currency i relative to the currency j, and $\mathbf{w} = (w_{12}, w_{13}, w_{14}, w_{23}, w_{24}, w_{34})$. Now assume that

$$\mathbf{w}_t = \mathbf{c} + \mathbf{Z}\mathbf{h}_t + \xi_t, \quad \xi_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\xi\xi}),$$
$$\mathbf{h}_{t+1} = \operatorname{diag}(\phi_1, \ldots, \phi_q)\mathbf{h}_t + \eta_t, \quad \eta_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}),$$

where $\mathbf{c}$ is a 6×1 mean vector, $\mathbf{\Sigma}_{\eta\eta}$ is diagonal, $\mathbf{h}_t = (h_{1t}, \ldots, h_{4t})'$ and h_{jt} is a latent factor for the jth currency at time t and

$$\mathbf{Z} = \begin{pmatrix} 1\,1\,0\,0 \\ 1\,0\,1\,0 \\ 1\,0\,0\,1 \\ 0\,1\,1\,0 \\ 0\,1\,0\,1 \\ 0\,0\,1\,1 \end{pmatrix}.$$

Since this is a linear Gaussian state-space model, the estimation of the parameters is straightforward by Kalman filtering methods.

Ray and Tsay (2000) introduced long-range dependence into the volatility factor model by supposing that $\mathbf{h}_t$ follows a fractionally integrated process such that

$$\mathbf{y}_t = \mathbf{V}_t^{1/2}\varepsilon_t, \quad \mathbf{V}_t^{1/2} = \mathrm{diag}(\exp(\mathbf{z}_1'\mathbf{h}_t/2), \ldots, \exp(\mathbf{z}_q'\mathbf{h}_t/2)),$$
$$(1-L)^d\mathbf{h}_t = \eta_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon}), \quad \eta_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}),$$

where $\mathbf{z}_i$ $(i = 1, \ldots, q)$ are $q \times 1$ vectors with $q < p$. In the fitting, the measurement equation is linearized as in Harvey et al. (1994).

Calvet et al. (2006) generalized the univariate Markov-switching multifractal (MSM) model proposed by Calvet and Fisher (2001) to the multivariate MSM and factor MSM models. The univariate model is given by

$$y_t = (M_{1,t}M_{2,t}\cdots M_{k,t})^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

where $M_{j,t}$ $(j \le k)$ are random volatility components, satisfying $\mathrm{E}(M_{j,t}) = 1$. Given $\mathbf{M}_t = (M_{1,t}, M_{2,t}, \ldots, M_{k,t})$, the stochastic volatility of return y_t is given by $\sigma^2 M_{1,t}M_{2,t}\cdots M_{k,t}$. Each $M_{j,t}$ follows a hidden Markov chain as follows:

$$M_{j,t} \text{ drawn from distribution } M, \text{with probability } \gamma_j,$$
$$M_{j,t} = M_{j,t-1}, \text{with probability } 1-\gamma_j,$$

where $\gamma_j = 1 - (1-\gamma)^{(b^{j-k})}$, $(0 < \gamma < 1, b > 1)$ and the distribution of M is binomial, giving values m or $2-m$ $(m \in [1,2])$ with equal probability. Thus, the MSM model is governed by four parameters (m, σ, b, γ), which are estimated by the maximum likelihood method.

For the bivariate MSM model, we consider the vector of the random volatility component $\mathbf{M}_{j,t} = (M_{j,t}^1, M_{j,t}^2)'$ $(j \le k)$. Then, the bivariate model is given by

$$\mathbf{y}_t = (\mathbf{M}_{1,t} \odot \mathbf{M}_{2,t} \odot \cdots \odot \mathbf{M}_{k,t})^{1/2} \odot \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_2(\mathbf{0}, V),$$

where $\odot$ denotes the element-by-element product. For each component $\mathbf{M}_{j,t}$ in the bivariate model, Calvet et al. (2006) assumed that volatility arrivals are correlated but not necessarily simultaneous. For details, let $s_{j,t}^i$ $(i = 1, 2)$ denote the random variable equal to 1 if there is an arrival on $M_{j,t}^i$ with prob-

ability γ_j, and equal to 0 otherwise. Thus, each $s^i_{j,t}$ follows the Bernoulli distribution. At this stage, Calvet et al. (2006) introduced the correlation coefficient λ, giving the conditional probability $P(s^2_{j,t} = 1|s^1_{j,t} = 1) = (1-\lambda)\gamma_j + \lambda$. They showed that arrivals are independent if $\lambda = 0$, and simultaneous if $\lambda = 1$. Given the realization of the arrival vectors $s^1_{j,t}$ and $s^2_{j,t}$, the construction of the volatility components $\mathbf{M}_{j,t}$ is based on a bivariate distribution $\mathbf{M} = (M_1, M_2)$. If arrivals hit both series ($s^1_{j,t} = s^2_{j,t} = 1$), the state vector $\mathbf{M}_{j,t}$ is drawn from $\mathbf{M}$. If only one series i ($i = 1, 2$) receives an arrival, the new component $M^i_{j,t}$ is sampled from the marginal M^i of the bivariate distribution $\mathbf{M}$. Finally, $\mathbf{M}_{j,t} = \mathbf{M}_{j,t-1}$ if there is no arrival ($s^1_{j,t} = s^2_{j,t} = 0$). They assumed that $\mathbf{M}$ has a bivariate binomial distribution controlled by m^1 and m^2, in parallel fashion to the univariate case. Again, the closed-form solution of the likelihood function is available. This approach can be extended to a general multivariate case. As the number of parameters therefore grows at least as fast as a quadratic function of p, Calvet et al. (2006) proposed not only the multivariate MSM model but also the factor MSM model.

The factor MSM model based on q volatility factors $\mathbf{f}^l_t = (f^l_{1,t}, \ldots, f^l_{k,t})'$, $(f^l_{j,t} > 0)$ $(l = 1, 2, \ldots, q)$ is given by

$$\begin{aligned}
\mathbf{y}_t &= \left(\mathbf{M}_{1,t} \odot \mathbf{M}_{2,t} \odot \cdots \odot \mathbf{M}_{k,t}\right)^{1/2} \odot \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_2(\mathbf{0}, V), \\
\mathbf{M}_{j,t} &= (M^1_{j,t}, M^2_{j,t}, \ldots, M^p_{j,t})', \quad (j \leq k), \\
M^i_{j,t} &= C_i \left(f^1_{j,t}\right)^{w^i_1} \left(f^2_{j,t}\right)^{w^i_2} \cdots \left(f^q_{j,t}\right)^{w^i_q} \left(u^i_{j,t}\right)^{w^i_{q+1}},
\end{aligned}$$

where the weights are nonnegative and add up to 1, and the constant C_i is chosen to guarantee that $\mathrm{E}(M^i_{j,t}) = 1$, and is thus not a free parameter. Calvet et al. (2006) specified the model as follows. For each vector f^l_t, $f^l_{j,t}$ follows a univariate MSM process with parameters (b, γ, m^l). The volatility of each asset i is also affected by an idiosyncratic shock $\mathbf{u}^i_t = (u^i_{1,t}, \ldots, u^i_{k,t})'$, which is specified by parameters (b, γ, m^{q+i}). Draws of the factors $f^l_{j,t}$ and idiosyncratic shocks $u^i_{j,t}$ are independent, but timing of arrivals may be correlated. Factors and idiosyncratic components thus follow the univariate MSM with identical frequencies.

3.2 Mean factor model

Another type of factor MSV model is considered in Pitt and Shephard (1999), who, following a model proposed in Kim et al. (1998), worked with the specification

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{V}_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}), \tag{18}$$

$$\mathbf{f}_t = \mathbf{D}_t^{1/2}\gamma_t, \quad \gamma_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}), \tag{19}$$

$$\mathbf{h}_{t+1} = \mu + \mathbf{\Phi}(\mathbf{h}_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}_{p+q}(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}), \tag{20}$$

where

$$\mathbf{V}_t = \text{diag}(\exp(h_{1t}), \ldots, \exp(h_{pt})), \tag{21}$$

$$\mathbf{D}_t = \text{diag}\left(\exp(h_{p+1,t}), \ldots, \exp(h_{p+q,t})\right), \tag{22}$$

$$\mathbf{\Phi} = \text{diag}(\phi_1, \ldots, \phi_{p+q}), \tag{23}$$

$$\mathbf{\Sigma}_{\eta\eta} = \text{diag}(\sigma_{1,\eta\eta}, ..., \sigma_{p+q,\eta\eta}) \tag{24}$$

and $\mathbf{h}_t = (h_{1t}, \ldots, h_{pt}, h_{p+1,t}, ..., h_{p+q,t})$. For identification purposes, the $p \times q$ loading matrix $\mathbf{B}$ is assumed to be such that $b_{ij} = 0$ for $(i < j,\ i \leq q)$ and $b_{ii} = 1$ $(i \leq q)$ with all other elements unrestricted. Thus, in this model, each of the factors and each of the errors evolve according to univariate SV models. A similar model was also considered by Jacquier et al. (1999) and Liesenfeld and Richard (2003) but under the restriction that $\mathbf{V}_t$ is not time-varying. Jacquier et al. (1999) estimated their model by MCMC methods, sampling h_{it} one at a time from its full conditional distribution, whereas Liesenfeld and Richard (2003) showed how the maximum likelihood estimation can be obtained by the efficient importance sampling method. For the more general model described above, Pitt and Shephard (1999) also employed a MCMC-based approach, now sampling $\mathbf{h}_t$ along the lines of Shephard and Pitt (1997). An even further generalization of this factor model was developed by Chib et al. (2006), who allowed for jumps in the observation model and a fat-tailed t distribution for the errors ε_t. The resulting model and its fitting are explained in Section 3.3. Alternative interesting approaches were also proposed by Diebold and Nerlove (1989) and King et al. (1994) in the framework of GARCH models using Kalman filter algorithms, but we omit the details to focus on the MSV models.

Lopes and Carvalho (2007) considered a general model which nests the models of Pitt and Shephard (1999) and Aguilar and West (2000), and extended it in two directions by (1) letting the matrix of factor loadings $\mathbf{B}$ be time-dependent and (2) allowing Markov switching in the common factor volatilities. The general model is given by (19)–(22), with

$$\mathbf{y}_t = \mathbf{B}_t\mathbf{f}_t + \mathbf{V}_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}),$$

$$\mathbf{h}_{t+1}^f = \mu_{s_t}^f + \mathbf{\Phi}^f\mathbf{h}_t^f + \eta_t^f, \quad \eta_t^f \sim \mathcal{N}_q\left(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}^f\right),$$

where $\mathbf{h}_t^f = (h_{p+1,t}, \ldots, h_{p+q,t})'$, $\mu^f = (\mu_{p+1}, \ldots, \mu_{p+q})'$, $\mathbf{\Phi}^f = \text{diag}(\phi_{p+1}, \ldots, \phi_{p+q})$ and $\mathbf{\Sigma}_{\eta\eta}^f$ is the nondiagonal covariance matrix. Letting the $pq - q(q+1)/2$ unconstrained elements of $\text{vec}(\mathbf{B}_t)$ be $\mathbf{b}_t = (b_{21,t}, b_{31,t}, \ldots, b_{pq,t})'$, they assumed that each element of $\mathbf{b}_t$ follows an AR(1) process. Following

So et al. (1998), where the fitting was based on the work of Albert and Chib (1993), they assumed μ_{s_t} followed a Markov switching model, where s_t follows a multistate first-order Markovian process. Lopes and Carvalho (2007) applied this model to two datasets: (1) returns on daily closing spot rates for six currencies relative to the US dollar (Deutschemark, British pound, Japanese yen, French franc, Canadian dollar, Spanish peseta), and (2) returns on daily closing rates for four Latin American stock markets indices. In the former application, they used $q = 3$ factors and in the latter case $q = 2$ factors.

Han (2006) modified the model of Pitt and Shephard (1999) and Chib et al. (2006) by allowing the factors to follow an AR(1) process:

$$\mathbf{f}_t = \mathbf{c} + \mathbf{A}\mathbf{f}_{t-1} + \mathbf{D}_t^{1/2}\gamma_t, \quad \gamma_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}). \tag{25}$$

The model was fit by adapting the approach of Chib et al. (2006) and applied to a collection of 36 arbitrarily chosen stocks to examine the performance of various portfolio strategies.

3.3 Bayesian analysis of mean factor MSV model

We describe the fitting of factor models in the context of the general model of Chib et al. (2006). The model is given by

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{K}_t\mathbf{q}_t + \mathbf{V}_t^{1/2}\mathbf{\Lambda}_t^{-1}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}), \tag{26}$$

where $\mathbf{\Lambda}_t = \text{diag}(\lambda_{1t}, \ldots, \lambda_{pt})$, $\mathbf{q}_t$ is p independent Bernoulli "jump" random variables and $\mathbf{K}_t = \text{diag}(k_{1t}, \ldots, k_{pt})$ are jump sizes. Assume that each element q_{jt} of $\mathbf{q}_t$ takes the value 1 with probability κ_j and the value 0 with probability $1 - \kappa_j$, and that each element u_{jt} of $\mathbf{u}_t = \mathbf{V}_t^{1/2}\mathbf{\Lambda}_t^{-1}\varepsilon_t$ follows an independent Student t distribution with degrees of freedom $\nu_j > 2$, which we express in hierarchical form as

$$u_{jt} = \lambda_{jt}^{-1/2}\exp(h_{jt}/2)\varepsilon_{jt}, \quad \lambda_{jt} \overset{i.i.d.}{\sim} \mathcal{G}\left(\frac{\nu_j}{2}, \frac{\nu_j}{2}\right), \quad t = 1, 2, \ldots, n. \tag{27}$$

The ε_t and $\mathbf{f}_t$ are assumed to be independent and

$$\begin{pmatrix} \mathbf{V}_t^{1/2}\varepsilon_t \\ \mathbf{f}_t \end{pmatrix} | \mathbf{V}_t, \mathbf{D}_t, \mathbf{K}_t, \mathbf{q}_t \sim \mathcal{N}_{p+q}\left\{\mathbf{0}, \begin{pmatrix} \mathbf{V}_t & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_t \end{pmatrix}\right\}$$

are conditionally independent Gaussian random vectors. The time-varying variance matrices $\mathbf{V}_t$ and $\mathbf{D}_t$ are defined by (20) and (21). Chib et al. (2006) assumed that the variables $\zeta_{jt} = \ln(1 + k_{jt})$, $j \leq p$, are distributed as $\mathcal{N}(-0.5\delta_j^2, \delta_j^2)$, where $\delta = (\delta_1, \ldots, \delta_p)'$ are unknown parameters.

We may calculate the number of parameters and latent variables as follows. Let β denote the free elements of $\mathbf{B}$ after imposing the identifying restrictions. Let $\mathbf{\Sigma}_{\eta\eta} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\mathbf{\Sigma}_{\eta\eta}^f = \text{diag}(\sigma_{p+1}^2, \ldots, \sigma_{p+q}^2)$. Then there are $pq - (q^2+q)/2$ elements in β. The model has $3(p+q)$ parameters $\theta_j = (\phi_j, \mu_j, \sigma_j)$ $(1 \le j \le p+q)$ in the autoregressive processes (20) of $\{h_{jt}\}$. We also have p degrees of freedom $\nu = (\nu_1, \ldots, \nu_p)$, p jump intensities $\kappa = (\kappa_1, \ldots, \kappa_p)$ and p jump variances $\delta = (\delta_1, \ldots, \delta_p)$. If we let $\psi = (\beta, \theta_1, \ldots, \theta_p, \nu, \delta, \kappa)$ denote the entire list of parameters, then the dimension of ψ is 688 when $p = 50$ and $q = 8$. Furthermore, the model contains $n(p+q)$ latent volatilities $\{\mathbf{h}_t\}$ that appears nonlinearly in the specification of $\mathbf{V}_t$ and $\mathbf{D}_t$, $2np$ latent variables $\{\mathbf{q}_t\}$ and $\{\mathbf{k}_t\}$ associated with the jump component and np scaling variables $\{\lambda_t\}$.

To conduct the prior-posterior analysis of this model, Chib et al. (2006) focused on the posterior distribution of the parameters and the latent variables:

$$\pi\left(\beta, \{\mathbf{f}_t\}, \{\theta_j\}, \{\mathbf{h}_{j.}\}, \{\nu_j\}, \{\lambda_{j.}\}, \{\delta_j\}, \{\kappa_j\}, \{\zeta_{j.}\}, \{\mathbf{q}_{j.}\} | Y_n\right), \qquad (28)$$

where the notation $\mathbf{z}_{j.}$ is used to denote the collection $(z_{j1}, \ldots, z_{jn})$. They sampled this distribution by MCMC methods through the following steps:

1. Sample β. The full conditional distribution of β is given by

$$\pi(\beta | Y_n, \{\mathbf{h}_{j.}\}, \{\zeta_{j.}\}, \{\mathbf{q}_{j.}\}, \{\lambda_{j.}\}) \propto p(\beta) \prod_{t=1}^{n} \mathcal{N}_p(\mathbf{y}_t | \mathbf{K}_t \mathbf{q}_t, \mathbf{\Omega}_t),$$

where $p(\beta)$ is the normal prior,

$$\mathbf{\Omega}_t = \mathbf{V}_t^* + \mathbf{B}\mathbf{D}_t\mathbf{B}' \quad \text{and} \quad \mathbf{V}_t^* = \mathbf{V}_t \odot \text{diag}(\lambda_{1t}^{-1}, \ldots, \lambda_{pt}^{-1}).$$

To sample from this density, Chib et al. (2006) employed the M-H algorithm (Chib and Greenberg (1995)), following Chib and Greenberg (1994) and taking the proposal density to be multivariate-t, $T(\beta | \mathbf{m}, \mathbf{\Sigma}, v)$, where $\mathbf{m}$ is the approximate mode of $l = \ln\{\prod_{t=1}^n \mathcal{N}_p(\mathbf{y}_t | \mathbf{K}_t\mathbf{q}_t, \mathbf{\Omega}_t)\}$, and $\mathbf{\Sigma}$ is minus the inverse of the second derivative matrix of l. Then, a proposal value β^* is drawn from $T(\mathbf{m}, \mathbf{\Sigma}, v)$ and accepted with probability

$$\alpha(\beta, \beta^* | \tilde{\mathbf{y}}, \{\mathbf{h}_{j.}\}, \{\lambda_{j.}\})$$
$$= \min\left\{1, \frac{p(\beta^*) \prod_{t=1}^n \mathcal{N}_p\left(\tilde{\mathbf{y}}_\mathbf{t} | \mathbf{0}, \mathbf{V}_t^* + \mathbf{B}^*\mathbf{D}_t\mathbf{B}^{*\prime}\right) T(\beta | \mathbf{m}, \mathbf{\Sigma}, v)}{p(\beta) \prod_{t=1}^n \mathcal{N}_p\left(\left(\tilde{\mathbf{y}}_t | \mathbf{0}, \mathbf{V}_t^* + \mathbf{B}\mathbf{D}_t\mathbf{B}'\right) T(\beta^* | \mathbf{m}, \mathbf{\Sigma}, v)\right.}\right\},$$

where β is the current value. If the proposal value is rejected, the next item of the chain is taken to be the current value β.

2. Sample $\{\mathbf{f}_t\}$. The distribution $\{\mathbf{f}_t\} | \tilde{Y}_n, \mathbf{B}, \mathbf{h}, \lambda$ can be divided into the product of the distributions $\mathbf{f}_t | \tilde{\mathbf{y}}_t, \mathbf{h}_t, \mathbf{h}_t^f, \lambda_t, \mathbf{B}$, which have Gaussian distribution with mean $\hat{\mathbf{f}}_t = \mathbf{F}_t\mathbf{B}'(\mathbf{V}_t^*)^{-1}\tilde{\mathbf{y}}_t$ and variance $\mathbf{F}_t = \left\{\mathbf{B}'(\mathbf{V}_t^*)^{-1}\mathbf{B} + \mathbf{D}_t^{-1}\right\}^{-1}$.

3. Sample $\{\theta_j\}$ and $\{\mathbf{h}_{j.}\}$. Given $\{\mathbf{f}_t\}$ and the conditional independence of the errors in (20), the model separates into q conditionally Gaussian state-space models. Let

$$z_{jt} = \begin{cases} \ln\{(y_{jt} - \alpha_{jt} - (\exp(\zeta_{jt}) - 1)q_{jt})^2 + c\} + \ln(\lambda_{jt}), & j \le p, \\ \ln(f_{j-p,t}^2 + c), & j \ge p+1, \end{cases}$$

where c is an "offset" constant that is set to 10^{-6}. Then from Kim et al. (1998) it follows that the $p+q$ state-space models can be subjected to an independent analysis for sampling the $\{\theta_j\}$ and $\{\mathbf{h}_{j.}\}$. In particular, the distribution of z_{jt}, which is h_{jt} plus a $\log \chi^2$ random variable with one degree of freedom, may be approximated closely by a seven-component mixture of normal distributions:

$$\begin{aligned} z_{jt}|s_{jt}, h_{jt} &\sim \mathcal{N}\left(h_{jt} + m_{s_{jt}}, v_{s_{jt}}^2\right), \\ h_{j,t+1} - \mu_j &= \phi_j\left(h_{j,t} - \mu_j\right) + \eta_{jt}, \quad j \le p+q, \end{aligned}$$

where s_{jt} is a discrete component indicator variable with mass function $\Pr(s_{jt} = i) = q_i$, $i \le 7$, $t \le n$, and $m_{s_{jt}}$, $v_{s_{jt}}^2$ and q_i are parameters that are reported in Chib et al. (2002). Thus, under this representation, conditioned on the transformed observations

$$p\left(\{\mathbf{s}_{j.}\}, \theta, \{\mathbf{h}_{j.}\}|\mathbf{z}\right) = \prod_{j=1}^{p+q} p\left(\mathbf{s}_{j.}, \theta_j, \mathbf{h}_{j.}|\mathbf{z}_{j.}\right),$$

which implies that the mixture indicators, log volatilities and series-specific parameters can be sampled series by series. Now, for each j, one can sample $(\mathbf{s}_{j.}, \theta_j, \mathbf{h}_{j.})$ by the univariate SV algorithm given by Chib et al. (2002). Briefly, $\mathbf{s}_{j.}$ is sampled from

$$p\left(\mathbf{s}_{j.}|\mathbf{z}_{j.}, \mathbf{h}_{j.}\right) = \prod_{t=1}^{n} p\left(s_{jt}|z_{jt}, h_{jt}\right),$$

where $p(s_{jt}|z_{jt}, h_{jt}) \propto p(s_{jt})\mathcal{N}\left(z_{jt}|\, h_{jt} + m_{s_{jt}}, v_{s_{jt}}^2\right)$ is a mass function with seven points of support. Next, θ_j is sampled by the M-H algorithm from the density $\pi(\theta_j|\mathbf{z}_{j.}, \mathbf{s}_{j.}) \propto p(\theta_j)p(\mathbf{z}_{j.}|\mathbf{s}_{j.}, \theta_j)$, where

$$p\left(\mathbf{z}_{j.}|\mathbf{s}_{j.}, \theta_j\right) = p\left(\mathbf{z}_{j1}|\mathbf{s}_{j.}, \theta_j\right) \prod_{t=2}^{n} p\left(\mathbf{z}_{jt}|\mathcal{F}_{j,t-1}^*, \mathbf{s}_{j.}, \theta_j\right) \tag{29}$$

and $p(z_{jt}|\mathcal{F}_{j,t-1}^*, \mathbf{s}_{j.}, \theta_j)$ is a normal density whose parameters are obtained by the Kalman filter recursions, adapted to the differing components, as indicated by the component vector $\mathbf{s}_{j.}$. Finally, $\mathbf{h}_{j.}$ is sampled

from $[\mathbf{h}_{j.}|\mathbf{z}_{j.}, \mathbf{s}_{j.}, \theta_j]$ by the simulation smoother algorithm of de Jong and Shephard (1995).

4. Sample $\{\nu_j\}$, $\{\mathbf{q}_{j.}\}$ and $\{\lambda_{j.}\}$. The degrees-of-freedom parameters, jump parameters and associated latent variables are sampled independently for each time series. The full conditional distribution of ν_j is given by

$$\begin{aligned} &\Pr(\nu_j|\mathbf{y}_{j.}, \mathbf{h}_j, \mathbf{B}, \mathbf{f}, \mathbf{q}_{j.}, \zeta_{j.}) \\ &\propto \Pr(\nu_j) \prod_{t=1}^{n} T(y_{jt}|\alpha_{jt} + \{\exp(\zeta_{jt}) - 1\} q_{jt}, \exp(h_{jt}), \nu_j), \end{aligned} \tag{30}$$

and one can apply the Metropolis-Hastings algorithm in a manner analogous to the case of β. Next, the jump indicators $\{\mathbf{q}_{j.}\}$ are sampled from the two-point discrete distribution,

$$\begin{aligned} &\Pr(q_{jt} = 1|\mathbf{y}_{j.}, \mathbf{h}_{j.}, \mathbf{B}, \mathbf{f}, \nu_j, \zeta_{j.}, \kappa_j) \\ &\propto \kappa_j T(y_{jt}|\alpha_{jt} + \{\exp(\zeta_{jt}) - 1\}, \exp(h_{jt}), \nu_j), \end{aligned}$$

$$\begin{aligned} &\Pr(q_{jt} = 0|\mathbf{y}_{j.}, \mathbf{h}_{j.}, \mathbf{B}, \mathbf{f}, \nu_j, \zeta_{j.}, \kappa_j) \\ &\propto (1 - \kappa_j) T(y_{jt}|\alpha_{jt}, \exp(h_{jt}), \nu_j), \end{aligned}$$

followed by the components of the vector $\{\lambda_{j.}\}$ from the density

$$\begin{aligned} &\lambda_{jt}|y_{jt}, h_{jt}, \mathbf{B}, \mathbf{f}, \nu_j, q_{jt}, \psi_{jt} \\ &\sim \mathcal{G}\left(\frac{\nu_j + 1}{2}, \frac{\nu_j + (y_{jt} - \alpha_{jt} - (\exp(\zeta_{jt}) - 1) q_{jt}))^2}{2\exp(h_{jt})}\right). \end{aligned}$$

5. Sample $\{\delta_j\}$ and $\{\zeta_{j.}\}$. For simulation efficiency reasons, δ_j and $\zeta_{j.}$ must also be sampled in one block. The full conditional distribution of δ_j is given by

$$\pi(\delta_j) \prod_{t=1}^{n} \mathrm{N}(\alpha_{jt} - 0.5\delta_j^2 q_{jt}, \delta_j^2 q_{jt}^2 + \exp(h_{jt})\lambda_{jt}^{-1}) \tag{31}$$

by the M-H algorithm. Once δ_j has been sampled, the vectors $\zeta_{j.}$ are sampled, bearing in mind that their posterior distribution is updated only when q_{jt} is 1. Therefore, when q_{jt} is 0, we sample ζ_{jt} from $\mathcal{N}(-0.5\delta_j^2, \delta_j^2)$, otherwise we sample from the distribution $\mathcal{N}(\Psi_{jt}(-0.5 + \exp(-h_{jt})\lambda_{jt} y_{jt}), \Psi_{jt})$, where $\Psi_{jt} = (\delta_j^{-2} + \exp(-h_{jt})\lambda_{jt})^{-1}$. The algorithm is completed by sampling the components of the vector κ independently from $\kappa_j|q_{j.} \sim \beta(u_{0j} + n_{1j}, u_{1j} + n_{0j})$, where n_{0j} is the count of $q_{jt} = 0$ and $n_{1j} = n - n_{0j}$ is the count of $q_{jt} = 1$.

A complete cycle through these various distributions completes one transition of our Markov chain. These steps are then repeated G times, where G is a large number, and the values beyond a suitable burn-in of say a 1,000 cycles are used for the purpose of summarizing the posterior distribution.

4 Dynamic Correlation MSV Model

Another way to model time-varying correlations is by constructing models that model the correlations (or functions of correlations) directly. We describe several such approaches in this section.

4.1 Modeling by reparameterization

One approach is illustrated by Yu and Meyer (2006) in the context of the bivariate SV model:

$$\begin{aligned}
\mathbf{y}_t &= \mathbf{V}_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon,t}), \quad \mathbf{\Sigma}_{\varepsilon\varepsilon,t} = \begin{pmatrix} 1 & \rho_t \\ \rho_t & 1 \end{pmatrix}, \\
\mathbf{h}_{t+1} &= \mu + \mathrm{diag}(\phi_1, \phi_2)(\mathbf{h}_t - \mu) + \eta_t, \quad \eta_t \sim \mathcal{N}_2\left(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2)\right), \\
q_{t+1} &= \psi_0 + \psi_1(q_t - \psi_0) + \sigma_\rho v_t, \quad v_t \sim \mathcal{N}(0,1), \\
\rho_t &= \frac{\exp(q_t) - 1}{\exp(q_t) + 1},
\end{aligned}$$

where $\mathbf{h}_0 = \mu$ and $q_0 = \psi_0$. The correlation coefficient ρ_t is then obtained from q_t by the Fisher transformation. Yu and Meyer (2006) estimated this model by MCMC methods with the help of the WinBUGS program and found that it was superior to other models, including the mean factor MSV model. However, the generalization of this bivariate model to higher dimensions is not easy because it is difficult to ensure the positive-definiteness of the correlation matrix $\mathbf{\Sigma}_{\varepsilon\varepsilon,t}$.

Another approach, introduced by Tsay (2005), is based on the Choleski decomposition of the time-varying correlation matrix. Specifically, one can consider the Choleski decomposition of the correlation matrix $\mathbf{\Sigma}_{\varepsilon\varepsilon,t}$ such that $\mathrm{Cov}(\mathbf{y}_t|\mathbf{h}_t) = \mathbf{L}_t\mathbf{V}_t\mathbf{L}_t'$. The outcome model is then given by $\mathbf{y}_t = \mathbf{L}_t\mathbf{V}_t^{1/2}\varepsilon_t$, $\varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$. As an example, when bivariate outcomes are involved we have

$$\mathbf{L}_t = \begin{pmatrix} 1 & 0 \\ q_t & 1 \end{pmatrix}, \quad \mathbf{V}_t = \mathrm{diag}(\exp(h_{1t}), \exp(h_{2t})).$$

Then,

$$
\begin{aligned}
y_{1t} &= \varepsilon_{1t} \exp(h_{1t}/2), \\
y_{2t} &= q_t \varepsilon_{1t} \exp(h_{1t}/2) + \varepsilon_{2t} \exp(h_{2t}/2),
\end{aligned}
$$

which shows that the distribution of $\mathbf{y}_t$ is modeled sequentially. We first let $y_{1t} \sim \mathcal{N}(0, \exp(h_{1t}))$ and then we let $y_{2t}|y_{1t} \sim \mathcal{N}(q_t y_{1t}, \exp(h_{2t}))$. Thus q_t is a slope of conditional mean and the correlation coefficient between y_{1t} and y_{2t} is given by

$$
\begin{aligned}
\mathrm{Var}(y_{1t}) &= \exp(h_{1t}), \\
\mathrm{Var}(y_{2t}) &= q_t^2 \exp(h_{1t}) + \exp(h_{2t}), \\
\mathrm{Cov}(y_{1t}, y_{2t}) &= q_t \exp(h_{1t}), \\
\mathrm{Corr}(y_{1t}, y_{2t}) &= \frac{q_t}{\sqrt{q_t^2 + \exp(h_{2t} - h_{1t})}}.
\end{aligned}
$$

As suggested in Asai et al. (2006), we let q_t follow an AR(1) process

$$
q_{t+1} = \psi_0 + \psi_1(q_t - \psi_0) + \sigma_\rho v_t, \quad v_t \sim \mathcal{N}(0, 1).
$$

The generalization to higher dimensions is straightforward. Let

$$
\mathbf{L}_t = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ q_{21,t} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ q_{p1,t} & \cdots & q_{p,p-1,t} & 1 \end{pmatrix}, \quad \mathbf{V}_t = \mathrm{diag}(\exp(h_{1t}), \ldots, \exp(h_{pt})),
$$

and

$$
\begin{aligned}
y_{1t} &= \varepsilon_{1t} \exp(h_{1t}/2), \\
y_{2t} &= q_{21,t} \varepsilon_{1t} \exp(h_{1t}/2) + \varepsilon_{2t} \exp(h_{2t}/2), \\
&\vdots \\
y_{pt} &= q_{p1,t} \varepsilon_{1t} \exp(h_{1t}/2) + \ldots + q_{p,p-1,t} \varepsilon_{p-1,t} \exp(h_{p-1,t}/2) + \varepsilon_{pt} \exp(h_{pt}/2),
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(y_{it}) &= \sum_{k=1}^{i} q_{ik,t}^2 \exp(h_{kt}), \quad q_{ii,t} \equiv 1, \quad i = 1, \ldots, p, \\
\mathrm{Cov}(y_{it}, y_{jt}) &= \sum_{k=1}^{i} q_{ik,t} q_{jk,t} \exp(h_{kt}), \quad i < j, \quad i = 1, \ldots, p-1,
\end{aligned}
$$

$$\mathrm{Corr}(y_{it}, y_{jt}) = \frac{\sum_{k=1}^{i} q_{ik,t} q_{jk,t} \exp(h_{kt})}{\sqrt{\sum_{k=1}^{i} q_{ik,t}^2 \exp(h_{kt}) \sum_{k=1}^{j} q_{jk,t}^2 \exp(h_{kt})}}, \quad i < j,$$

where q_{it} now follows the AR(1) process:

$$q_{i,t+1} = \psi_{i,0} + \psi_{i,1}(q_{i,t} - \psi_0) + \sigma_{i,\rho} v_{it}, \quad v_{it} \sim \mathcal{N}(0,1).$$

Jungbacker and Koopman (2006) considered a similar model with $\mathbf{L}_t = \mathbf{L}$ and estimated the parameters of the model by the Monte Carlo likelihood method. As in the one-factor case, they used the data set for the daily exchange rate returns of the British pound, the Deutschemark and the Japanese yen against the US dollar.

4.2 Matrix exponential transformation

For any $p \times p$ matrix $\mathbf{A}$, the matrix exponential transformation is defined by the following power-series expansion:

$$\exp(\mathbf{A}) \equiv \sum_{s=0}^{\infty} \frac{1}{s!} \mathbf{A}^s,$$

where $\mathbf{A}^0$ is equal to a $p \times p$ identity matrix. For any real positive-definite matrix $\mathbf{C}$, there exists a real symmetric $p \times p$ matrix $\mathbf{A}$ such that

$$\mathbf{C} = \exp(\mathbf{A}).$$

Conversely, for any real symmetric matrix $\mathbf{A}$, $\mathbf{C} = \exp(\mathbf{A})$ is a positive-definite matrix; see Lemma 1 of Chiu et al. (1996) and Kawakatsu (2006). If $\mathbf{A}_t$ is a $p \times p$ real symmetric matrix, there exists a $p \times p$ orthogonal matrix $\mathbf{B}_t$ and a $p \times p$ real diagonal matrix $\mathbf{H}_t$ of eigenvalues of $\mathbf{A}$ such that $\mathbf{A}_t = \mathbf{B}_t \mathbf{H}_t \mathbf{B}_t'$ and

$$\exp(\mathbf{A}_t) = \mathbf{B}_t \left(\sum_{s=0}^{\infty} \frac{1}{s!} \mathbf{H}_t^s \right) \mathbf{B}_t' = \mathbf{B}_t \exp(\mathbf{H}_t) \mathbf{B}_t' \ .$$

Thus, we consider the matrix exponential transformation for the covariance matrix $\mathrm{Var}(\mathbf{y}_t) = \mathbf{\Sigma}_t = \exp(\mathbf{A}_t)$, where $\mathbf{A}_t$ is a $p \times p$ real symmetric matrix such that $\mathbf{A}_t = \mathbf{B}_t \mathbf{H}_t \mathbf{B}_t'$ ($\mathbf{H}_t = \mathrm{diag}(h_{1t}, \ldots, h_{pt})$). Note that

$$\boldsymbol{\Sigma}_t = \mathbf{B}_t \mathbf{V}_t \mathbf{B}_t', \quad \mathbf{V}_t = \operatorname{diag}(\exp(h_{1t}), \ldots, \exp(h_{pt})),$$
$$\boldsymbol{\Sigma}_t^{-1} = \mathbf{B}_t' \mathbf{V}_t^{-1} \mathbf{B}_t, \quad |\boldsymbol{\Sigma}_t| = \exp\left(\sum_{i=1}^{p} h_{it}\right).$$

We model the dynamic structure of covariance matrices through $\alpha_t = \operatorname{vech}(\mathbf{A}_t)$. We may consider a first-order autoregressive process for α_t,

$$\begin{aligned} \mathbf{y}_t | \mathbf{A}_t &\sim \mathcal{N}_p(\mathbf{0}, \exp(\mathbf{A}_t)), \\ \alpha_{t+1} &= \mu + \boldsymbol{\Phi}(\alpha_t - \mu) + \eta_t, \quad (\boldsymbol{\Phi} : \text{diagonal}), \\ \alpha_t &= \operatorname{vech}(\mathbf{A}_t), \quad \eta_t \sim \mathcal{N}_{p(p+1)/2}(\mathbf{0}, \boldsymbol{\Sigma}_{\eta\eta}), \end{aligned}$$

as suggested in Asai et al. (2006). The estimation of this model can be done using a MCMC or a simulated maximum likelihood estimation, but it is not straightforward to interpret the parameters.

4.3 Wishart process

4.3.1 Standard model

Another way to obtain a time-varying correlation matrix is by the approach of Philipov and Glickman (2006a, 2006b), who assumed that the conditional covariance matrix $\boldsymbol{\Sigma}_t$ follows an inverted Wishart distribution with parameters that depend on the past covariance matrix $\boldsymbol{\Sigma}_{t-1}$. In particular,

$$\begin{aligned} \mathbf{y}_t | \boldsymbol{\Sigma}_t &\sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_t), \\ \boldsymbol{\Sigma}_t | \nu, \mathbf{S}_{t-1} &\sim \mathcal{IW}_p(\nu, \mathbf{S}_{t-1}), \end{aligned}$$

where $\mathcal{IW}(\nu_0, \mathbf{Q}_0)$ denotes an inverted Wishart distribution with parameters $(\nu_0, \mathbf{Q}_0)$,

$$\begin{aligned} \mathbf{S}_{t-1} &= \frac{1}{\nu} \mathbf{A}^{1/2} \left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^d \mathbf{A}^{1/2\prime}, \\ \mathbf{A} &= \mathbf{A}^{1/2} \mathbf{A}^{1/2\prime}, \end{aligned} \tag{32}$$

and $\mathbf{A}^{1/2}$ is a Choleski decomposition of a positive-definite symmetric matrix $\mathbf{A}$ and $-1 < d < 1$. Asai and McAleer (2007) pointed out that it is also possible to parameterize $\mathbf{S}_{t-1}$ as $\nu^{-1} \left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^{d/2} \mathbf{A} \left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^{d/2\prime}$.

The conditional expected values of $\boldsymbol{\Sigma}_t^{-1}$ and $\boldsymbol{\Sigma}_t$ are

$$\mathrm{E}\left(\boldsymbol{\Sigma}_t^{-1}|\nu, \mathbf{S}_{t-1}\right) = \nu\mathbf{S}_{t-1} = \mathbf{A}^{1/2}\left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^d \mathbf{A}^{1/2\prime},$$
$$\mathrm{E}\left(\boldsymbol{\Sigma}_t|\nu, \mathbf{S}_{t-1}\right) = \frac{1}{\nu-p-1}\mathbf{S}_{t-1}^{-1} = \frac{\nu}{\nu-p-1}\mathbf{A}^{-1/2}\left(\boldsymbol{\Sigma}_{t-1}\right)^d \mathbf{A}^{-1/2\prime},$$

respectively. Thus, the scale parameter d expresses the overall strength of the serial persistence in the covariance matrix over time. On the basis of the process of the logarithm of the determinant and asymptotic behavior of expectation of the determinant, they assumed that $|d| < 1$, although it is natural to assume that $0 < d < 1$. Notice that when $d = 0$, for example, the serial persistence disappears and we get

$$\mathrm{E}\left(\boldsymbol{\Sigma}_t^{-1}|\nu, \mathbf{S}_{t-1}\right) = \mathbf{A},$$
$$\mathrm{E}\left(\boldsymbol{\Sigma}_t|\nu, \mathbf{S}_{t-1}\right) = \frac{\nu}{\nu-p-1}\mathbf{A}^{-1}.$$

The matrix $\mathbf{A}$ in this model is a measure of the intertemporal sensitivity and determines how the elements of the current period covariance matrix $\boldsymbol{\Sigma}_t$ are related to the elements of the previous period covariance matrix. When $\mathbf{A} = \mathbf{I}$, we note that

$$\mathrm{E}\left(\boldsymbol{\Sigma}_t^{-1}|\nu, \mathbf{S}_{t-1}\right) = \begin{cases} \boldsymbol{\Sigma}_{t-1}^{-1}, & d = 1, \\ \mathbf{I}, & d = 0, \\ \boldsymbol{\Sigma}_{t-1}, & d = -1. \end{cases}$$

Philipov and Glickman (2006b) estimated this model from a Bayesian approach and proposed a MCMC algorithm to estimate their models using monthly return data of five industry portfolios (manufacturing, utilities, retail/wholesale, financial and other) in NYSE, AMEX and NASDAQ stocks. Under the prior

$$\mathbf{A} \sim \mathcal{IW}_p(\nu_0, \mathbf{Q}_0), \quad d \sim \pi(d), \quad \nu - p \sim \mathcal{G}(\alpha, \beta)$$

with $\boldsymbol{\Sigma}_0$ assumed known, the MCMC algorithm is implemented as follows:

1. Sample $\boldsymbol{\Sigma}_t|\{\boldsymbol{\Sigma}_s\}_{s\neq t}, \mathbf{A}, \nu, d, Y_n$ $(t = 1, \ldots, n-1)$, where $Y_n = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Given a current sampler $\boldsymbol{\Sigma}_t$, we generate a candidate $\boldsymbol{\Sigma}_t^* \sim \mathcal{W}_p(\tilde{\nu}, \tilde{\mathbf{S}}_{t-1})$, where $\mathcal{W}_p(\tilde{\nu}, \tilde{\mathbf{S}}_{t-1})$ denotes a Wishart distribution with parameters $(\tilde{\nu}, \tilde{\mathbf{S}}_{t-1})$,

$$\tilde{\nu} = \nu(1-d) + 1,$$
$$\tilde{\mathbf{S}}_{t-1} = \mathbf{S}_{t-1}^{-1} + \mathbf{y}_t\mathbf{y}_t',$$
$$\mathbf{S}_{t-1} = \frac{1}{\nu}(\mathbf{A}^{1/2})\left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^d(\mathbf{A}^{1/2})',$$

and accept it with probability

$$\min\left\{\frac{|\boldsymbol{\Sigma}_t^*|^{(\nu d-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left\{\nu\mathbf{A}^{-1}\left(\boldsymbol{\Sigma}_t^*\right)^{-d}\boldsymbol{\Sigma}_{t+1}^{-1}\right\}\right]}{|\boldsymbol{\Sigma}_t|^{(\nu d-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left\{\nu\mathbf{A}^{-1}\left(\boldsymbol{\Sigma}_t\right)^{-d}\boldsymbol{\Sigma}_{t+1}^{-1}\right\}\right]},1\right\}.$$

2. Sample $\boldsymbol{\Sigma}_n|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n-1},\mathbf{A},\nu,d,Y_n\sim\mathcal{W}_p(\tilde{\nu},\tilde{\mathbf{S}}_{n-1})$.
3. Sample $\mathbf{A}|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\nu,d,\mathbf{y}\sim\mathcal{IW}_p(\tilde{\gamma},\tilde{\mathbf{Q}})$, where $\tilde{\gamma}=n\nu+\nu_0$, and

$$\tilde{\mathbf{Q}}^{-1}=\nu\left\{\sum_{t=1}^{n}\left(\boldsymbol{\Sigma}_t^{-1}\right)^{-d/2}\boldsymbol{\Sigma}_t^{-1}\left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^{-d/2}\right\}+\mathbf{Q}_0^{-1}.$$

4. Sample d from

$$\begin{aligned}&\pi(d|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\mathbf{A},\nu,\mathbf{y})\\&\propto\pi(d)\exp\left[\frac{\nu d}{2}\sum_{t=1}^{n}\log|\boldsymbol{\Sigma}_t|-\frac{1}{2}\sum_{t=1}^{n}\mathrm{tr}\left\{\mathbf{S}_t^{-1}\left(\boldsymbol{\Sigma}_{t-1}^{-1}\right)^{-d}\right\}\right].\end{aligned}$$

To sample d, Philipov and Glickman (2006b) suggested discretizing the conditional distribution; see Appendix A.2 of Philipov and Glickman (2006b). Alternatively, we may conduct an independent M-H algorithm using a candidate from a truncated normal distribution $\mathcal{TN}_{(0,1)}(\hat{d},\hat{V}_d)$, where $\mathcal{TN}_{(a,b)}(\mu,\sigma^2)$ denote a normal distribution with mean μ and variance σ^2 truncated on the interval (a,b), $\hat{d}$ is a mode of conditional posterior probability density $\pi(d|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\mathbf{A},\nu,\mathbf{y})$ and

$$\hat{V}_d=\left\{-\left.\frac{\partial^2\log\pi(d|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\mathbf{A},\nu,Y_n)}{\partial d^2}\right|_{d=\hat{d}}\right\}^{-1}.$$

5. Sample ν from

$$\begin{aligned}&\pi(\nu|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\mathbf{A},d,\mathbf{y})\\&\propto(\nu-p)^{\alpha-1}\exp\{-\beta(\nu-p)\}\left\{\frac{|\nu\mathbf{A}^{-1}|^{\nu/2}}{2^{\nu p}\prod_{j=1}^{p}\Gamma(\frac{\nu+j-1}{2})}\right\}^{n}\\&\qquad\times\exp\left[-\frac{\nu}{2}\sum_{t=1}^{n}\left\{\log|\mathbf{Q}_t|+\mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{Q}_t^{-1}\right)\right\}\right].\end{aligned}$$

As in the previous step, we may discretize the conditional distribution or conduct an independent M-H algorithm using a candidate from a truncated normal distribution $\mathcal{TN}_{(p,\infty)}(\hat{\nu},\hat{V}_\nu)$, where $\hat{\nu}$ is a mode of conditional posterior probability density $\pi(\nu|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\mathbf{A},d,\mathbf{y})$ and

$$\hat{V}_\nu=\left\{-\left.\frac{\partial^2\log\pi(\nu|\{\boldsymbol{\Sigma}_t\}_{t=1}^{n},\mathbf{A},d,Y_n)}{\partial\nu^2}\right|_{\nu=\hat{\nu}}\right\}^{-1}.$$

Asai and McAleer (2007) proposed two further models that are especially useful in higher dimensions. Let $\mathbf{Q}_t$ be a sequence of positive-definite matrices, which is used to define the correlation matrix $\mathbf{\Sigma}_{\varepsilon\varepsilon,t} = \mathbf{Q}_t^{*-1/2}\mathbf{Q}_t\mathbf{Q}_t^{*-1/2}$, where $\mathbf{Q}_t^*$ is a diagonal matrix whose (i,i)th element is the same as that of $\mathbf{Q}_t$. Then the first of their dynamic correlation (DC) MSV models is given by

$$\begin{aligned}
\mathbf{y}_t &= \mathbf{V}_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon\varepsilon,t}), \quad \mathbf{\Sigma}_{\varepsilon\varepsilon,t} = \mathbf{Q}_t^{*-1/2}\mathbf{Q}_t\mathbf{Q}_t^{*-1/2}, \\
\mathbf{h}_{t+1} &= \tilde{\mu} + \mathbf{\Phi}\mathbf{h}_t + \eta_t, \quad \eta_t \sim \mathcal{N}_p\left(\mathbf{0}, \mathbf{\Sigma}_{\eta\eta}\right), \quad (\mathbf{\Phi} \text{ and } \mathbf{\Sigma}_{\eta\eta} : \text{diagonal}) \\
\mathbf{Q}_{t+1} &= (1-\psi)\bar{\mathbf{Q}} + \psi\mathbf{Q}_t + \mathbf{\Xi}_t, \quad \mathbf{\Xi}_t \sim \mathcal{W}_p(\nu, \mathbf{\Lambda}).
\end{aligned}$$

Thus, in this model the MSV shocks are assumed to follow a Wishart process, where $\mathcal{W}_p(\nu, \mathbf{\Lambda})$ denotes a Wishart distribution with degrees-of-freedom parameter ν and scale matrix $\mathbf{\Lambda}$. The model guarantees that $\mathbf{P}_t$ is symmetric positive-definite under the assumption that $\bar{\mathbf{Q}}$ is positive-definite and $|\psi| < 1$. It is possible to consider a generalization of the model by letting $\mathbf{Q}_{t+1} = (\mathbf{1}\mathbf{1}' - \mathbf{\Psi}) \odot \bar{\mathbf{Q}} + \mathbf{\Psi} \odot \mathbf{Q}_t + \mathbf{\Xi}_t$, which corresponds to a generalization of the dynamic conditional correlation (DCC) model of Engle (2002).

The second DC MSV model is given by

$$\mathbf{Q}_{t+1}|\nu, \mathbf{S}_t \sim \mathcal{IW}_p(\nu, \mathbf{S}_t), \quad \mathbf{S}_t = \frac{1}{\nu}\mathbf{Q}_t^{-d/2}\mathbf{A}\mathbf{Q}_t^{-d/2},$$

where ν and $\mathbf{S}_t$ are the degrees of freedom and the time-dependent scale parameter of the Wishart distribution, respectively, $\mathbf{A}$ is a positive-definite symmetric parameter matrix, d is a scalar parameter and $\mathbf{Q}_t^{-d/2}$ is defined by using a singular value decomposition. The quadratic expression, together with $\nu \geq p$, ensures that the covariance matrix is symmetric and positive-definite. For convenience, it is assumed that $\mathbf{Q}_0 = \mathbf{I}_p$. Although their model is closely related to the models of Philipov and Glickman (2006a, 2006b), the MCMC fitting procedures are different. Asai and McAleer (2007) estimated these models using returns of the Nikkei 225 Index, the Hang Seng Index and the Straits Times Index.

Gourieroux et al. (2004) and Gourieroux (2006) used an alternative approach and derived a Wishart autoregressive process. Let $\mathbf{Y}_t$ and $\mathbf{\Gamma}$ denote, respectively, a stochastic symmetric positive-definite matrix of dimension $p \times p$ and a deterministic symmetric matrix of dimension $p \times p$. A Wishart autoregressive process of order 1 is defined to be a matrix process (denoted by *WAR(1)* process) with a conditional Laplace transform:

$$\Psi_t(\mathbf{\Gamma}) = \mathrm{E}_t\left[\exp\{\mathrm{tr}(\mathbf{\Gamma}\mathbf{Y}_{t+1})\}\right] = \frac{\exp\left[\mathrm{tr}\left\{\mathbf{M}'^{-1}\mathbf{M}\mathbf{Y}_t\right\}\right]}{|\mathbf{I} - 2\mathbf{\Sigma}\mathbf{\Gamma}|^{k/2}}, \tag{33}$$

where k is a scalar degree of freedom ($k < p-1$), $\mathbf{M}$ is a $p \times p$ matrix of autoregressive parameters and $\mathbf{\Sigma}$ is a $p \times p$ symmetric and positive-definite

matrix such that the maximal eigenvalue of $2\boldsymbol{\Sigma}\boldsymbol{\Gamma}$ is less than 1. Here E_t denotes the expectation conditional on $\{\mathbf{Y}_t, \mathbf{Y}_{t-1}, \dots, \}$. It can be shown that

$$\mathbf{Y}_{t+1} = \mathbf{M}\mathbf{Y}_t\mathbf{M}' + k\boldsymbol{\Sigma} + \eta_{t+1},$$

where $\mathrm{E}(\eta_{t+1}) = \mathbf{O}$. The conditional probability density function of $\mathbf{Y}_{t+1}$ is given by

$$f(\mathbf{Y}_{t+1}|\mathbf{Y}_t) = \frac{|\mathbf{Y}_{t+1}|^{(k-p-1)/2}}{2^{kp/2}\Gamma_p(k/2)|\boldsymbol{\Sigma}|^{k/2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_{t+1} + \mathbf{M}\mathbf{Y}_t\mathbf{M}')\right\}\right] \times_0 F_1(k/2; (1/4)\mathbf{M}\mathbf{Y}_t\mathbf{M}'\mathbf{Y}_{t+1}),$$

where Γ_p is the multidimensional gamma function and ${}_0F_1$ is the hypergeometric function of matrix augment; see Gourieroux et al. (2004) for details. When K is an integer and $\mathbf{Y}_t$ is a sum of outer products of k independent vector AR(1) processes such that

$$\mathbf{Y}_t = \sum_{j=1}^{k} \mathbf{x}_{jt}\mathbf{x}_{jt}', \tag{34}$$
$$\mathbf{x}_{jt} = \mathbf{M}\mathbf{x}_{j,t-1} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}),$$

we obtain the Laplace transform $\Psi_t(\boldsymbol{\Gamma})$ given by (33). Gourieroux et al. (2004) also introduced a Wishart autoregressive process of higher order. They estimated the WAR(1) using a series of intraday historical volatility–covolatility matrices for three stocks traded on the Toronto Stock Exchange. Finally, Gourieroux (2006) introduced the continuous-time Wishart process as the multivariate extension of the Cox–Ingersoll–Ross (CIR) model in Cox et al. (1985).

4.3.2 Factor model

Philipov and Glickman (2006a) proposed an alternative factor MSV model that assumes that the factor volatilities follow an unconstrained Wishart random process. Their model has close ties to the model in Philipov and Glickman (2006b), and is given by

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{V}^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}),$$
$$\mathbf{f}_t|\boldsymbol{\Sigma}_t \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_t), \quad \boldsymbol{\Sigma}_t|\nu, \mathbf{S}_{t-1} \sim \mathcal{IW}_q(\nu, \mathbf{S}_{t-1}),$$

where $\mathbf{S}_{t-1}$ is defined by (32). In other words, the conditional covariance matrix $\boldsymbol{\Sigma}_t$ of the factor $\mathbf{f}_t$ follows an inverse Wishart distribution whose parameter depends on the past covariance matrix $\boldsymbol{\Sigma}_{t-1}$. They implemented the model with $q = 2$ factors on return series data of 88 individual companies from the S&P500.

In another development, Carvalho and West (2006) proposed dynamic matrix-variate graphical models, which are based on dynamic linear models accommodated with the hyperinverse Wishart distribution that arises in the study of graphical models (Dawid and Lauritzen (1993), Carvalho and West (2006)). The starting point is the dynamic linear model

$$\begin{aligned}\mathbf{y}_t' &= \mathbf{X}_t'\boldsymbol{\Theta}_t + \mathbf{u}_t', \quad \mathbf{u}_t \sim \mathcal{N}_p(\mathbf{0}, v_t\boldsymbol{\Sigma}),\\ \boldsymbol{\Theta}_t &= \mathbf{G}_t\boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t, \quad \boldsymbol{\Omega}_t \sim \mathcal{N}_{q\times p}(O, \mathbf{W}_t, \boldsymbol{\Sigma}),\end{aligned}$$

where $\mathbf{y}_t$ is the $p \times 1$ vector of observations, $\mathbf{X}_t$ is a known $q \times 1$ vector of explanatory variables, $\boldsymbol{\Theta}_t$ is the $q \times p$ matrix of states, $\mathbf{u}_t$ is the $p \times 1$ innovation vector for observation, $\boldsymbol{\Omega}_t$ is the $q \times p$ innovation matrix for states, $\mathbf{G}_t$ is a known $q \times q$ matrix and $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix. $\boldsymbol{\Omega}_t$ follows a matrix-variate normal with mean $\mathbf{O}$ $(q \times p)$, left covariance matrix $\mathbf{W}_t$ and right covariance matrix $\boldsymbol{\Sigma}$; in other words, any column ω_{it} of $\boldsymbol{\Omega}_t$ has a multivariate normal distribution $\mathcal{N}_q(\mathbf{0}, \sigma_{ii}\mathbf{W}_t)$, while any row ω_t^i of $\boldsymbol{\Omega}_t$, $\omega_t^{i\prime}$ has a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, w_{ii,t}\boldsymbol{\Sigma})$. Next, we suppose that $\boldsymbol{\Sigma} \sim \mathcal{HIW}_p(b, \mathbf{D})$, the hyperinverse Wishart distribution with a degree-of-freedom parameter b and location matrix $\mathbf{D}$. It should be noted that the dynamic linear model with $\boldsymbol{\Sigma} \sim \mathcal{HIW}_p(b, \mathbf{D})$ can be handled from the Bayesian perspective without employing simulation-based techniques. Finally, instead of time-invariant $\boldsymbol{\Sigma}$, Carvalho and West (2006) suggested a time-varying process given by

$$\begin{aligned}\boldsymbol{\Sigma}_t &\sim \mathcal{HIW}_p(b_t, \mathbf{S}_t),\\ b_t &= \delta b_{t-1} + 1,\\ \mathbf{S}_t &= \delta\mathbf{S}_{t-1} + \mathbf{v}_t\mathbf{v}_t',\end{aligned}$$

where $\mathbf{v}_t$ is defined by Theorem 1 of Carvalho and West (2006). Intuitively, $\mathbf{v}_t$ is the residual from the observation equation. As $\boldsymbol{\Sigma}_t$ appears in both of the observation and state equations, the proposed dynamic matrix-variate graphical model can be considered as a variation of the "factor MSV model with MSV error." Setting $\delta = 0.97$, Carvalho and West (2006) applied the dynamic matrix-variate graphical models to two datasets; namely, (1) 11 international currency exchange rates relative to the US dollar and (2) 346 securities from the S&P500 stock index.

5 Conclusion

We have conducted a comprehensive survey of the major current themes in the formulation of MSV models. In time, further significant developments can be expected, perhaps fostered by the overview and details delineated in this paper, especially in the fitting of high-dimensional models. Open problems

remain, primarily in the modeling of leverage effects, especially in relation to general specifications of cross-leverage effects embedded within multivariate heavy-tailed or skewed error distributions. We also expect that interest in the class of factor-based MSV models and DC models will grow as these approaches have shown promise in the modeling of high-dimensional data.

References

Aas, K. and Haff, I. H. (2006): The generalized hyperbolic skew Student's t-distribution. *Journal of Financial Econometrics* **4**, 275–309.

Aguilar, O. and West, M. (2000): Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics* **18**, 338–357.

Albert, J. H. and Chib, S. (1993): Bayesian inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics* **11**, 1–15.

Andersen, T., Bollerslev, T., Diebold, F. X. and Labys, P. (2003): Modeling and forecasting realized volatility. *Econometrica* **71**, 579–625.

Arellano-Valle, R. B. and Azzalini, A. (2006): On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**, 561–574.

Asai, M. and McAleer, M. (2006): Asymmetric multivariate stochastic volatility. *Econometric Reviews* **25**, 453–473.

Asai, M. and McAleer, M. (2007): The structure of dynamic correlations in multivariate stochastic volatility models. *Unpublished paper: Faculty of Economics, Soka University*.

Asai, M., McAleer, M. and Yu, J. (2006). Multivariate stochastic volatility: A review. *Econometric Reviews* **25**, 145–175.

Azzalini, A. (2005): The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**, 159–188.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perterbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society Series B* **65**, 367–389.

Barndorff-Nielsen, O. E. (1977): Exponentially decreasing distributions for the logarithm of the particle size. *Proceedings of the Royal Society London Series A Mathematical and Physical Sciences* **353**, 401–419.

Barndorff-Nielsen, O. E. and Shephard, N. (2001): Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society Series B* **63**, 167–241.

Barndorff-Nielsen, O. E. and Shephard, N. (2004): Econometric analysis of realised covariation: High frequency based covariance, regression and correlation in financial economics. *Econometrica* **72**, 885–925.

Bauwens, L. and Laurent, S. (2005): A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *Journal of Business and Economic Statistics* **23**, 346–354.

Bauwens, L., Laurent, S. and Rombouts, J. V. K. (2006): Multivariate GARCH: A survey. *Journal of Applied Econometrics* **21**, 79–109.

Bollerslev, T. (1990): Modelling the coherence in the short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* **72**, 498–505.

Bollerslev, T., Engle, R. F. and Woodridge, J. (1988): A capital asset pricing model with time varying covariances. *Journal of Political Economy* **96**, 116–131.

Bos, C. S. and Shephard, N. (2006): Inference for adaptive time series models: Stochastic volatility and conditionally Gaussian state space form. *Econometric Reviews* **25**, 219–244.

Broto, C. and Ruiz, E. (2004): Estimation methods for stochastic volatility models: A survey. *Journal of Economic Survey* **18**, 613–649.

Calvet, L. E. and Fisher, A. J. (2001): Forecasting multifractal volatility. *Journal of Econometrics* **105**, 27–58.

Calvet, L. E., Fisher, A. J. and Thompson, S. B. (2006): Volatility comovement: A multifrequency approach. *Journal of Econometrics* **131**, 179–215.

Carvalho, C. M. and West, M. (2006): Dynamic matrix-variate graphical models. *Bayesian Analysis* **1**, 1–29.

Chan, D., Kohn, R. and Kirby, C. (2006): Multivariate stochastic volatility models with correlated errors. *Econometric Reviews* **25**, 245–274.

Chib, S. (2001): Markov chain Monte Carlo methods: Computation and inference. *In: Heckman, J. J. and Leamer, E. (Eds.): Handbook of Econometrics* **5**, 3569–3649. North-Holland, Amsterdam.

Chib, S. and Greenberg, E. (1994): Bayes inference for regression models with ARMA(p, q) errors. *Journal of Econometrics* **64**, 183–206.

Chib, S. and Greenberg, E. (1995): Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327–335.

Chib, S. and Greenberg, E. (1996): Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory* **12**, 409–431.

Chib, S. and Greenberg, E. (1998): Analysis of multivariate Probit models. *Biometrika* **85**, 347–361.

Chib, S., Nardari, F. and Shephard, N. (2002): Markov chain Monte Carlo methods for generalized stochastic volatility models. *Journal of Econometrics* **108**, 281–316.

Chib, S., Nardari, F. and Shephard, N. (2006): Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics* **134**, 341–371.

Chiu, T., Leonard, T. and Tsui, K. (1996): The matrix-logarithmic covariance model. *Journal of the American Statistical Association* **91**, 198–210.

Cox, J., Ingersoll, J. and Ross, S. (1985): A theory of the term structure of interest rates. *Econometrica* **53**, 385–407.

Dagpunar, J. S. (1989): An easily implemented generalized inverse Gaussian generator. *Communications in Statistics Simulations* **18**, 703–710.

Daníelsson, J. (1994): Stochastic volatility in asset prices: Estimation with simulated maximum likelihood. *Journal of Econometrics* **64**, 375–400.

Daníelsson, J. (1998): Multivariate stochastic volatility models: Estimation and a comparison with VGARCH models. *Journal of Empirical Finance* **5**, 155–173.

Dawid, A. P. and Lauritzen, S. L. (1993): Hyper-Markov laws in the statistical analysis. *Annals of Statistics* **3**, 1272–1317.

de Jong, P. and Shephard, N. (1995): The simulation smoother for time series models. *Biometrika* **82**, 339–350.

Dey, D. and Liu, J. (2005): A new construction for skew multivariate distributions. *Journal of Multivariate Analysis* **95**, 323–344.

Diebold, F. X. and Nerlove, M. (1989): The dynamics of exchange rate volatility: A multivariate latent-factor ARCH model. *Journal of Applied Econometrics* **4**, 1–22.

Doornik, J. A. (2002): *Object-Oriented Matrix Programming Using Ox* (3rd ed.). Timberlake Consultants Press, London. http://www.nuff.ox.ac.uk/Users/Doornik.

Durbin, J. and Koopman, S. J. (2002): A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**, 603–616.

Engle, R. F. (2002): Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* **20**, 339–350.

Engle, R. F. and Kroner, K. F. (1995): Multivariate simultaneous generalized ARCH. *Econometric Theory* **11**, 122–150.

Ferreira, J. T. A. S. and Steel, M. F. J. (2004): Bayesian multivariate regression analysis with a new class of skewed distributions. *Statistics Research Report 419, University of Warwick*.

Ghysels, E., Harvey, A. C. and Renault, E. (1996): Stochastic volatility. *In: G. S. M. Rao, C. R. (Ed.): Statistical Models in Finance (Handbook of Statistics)*, 119–191. North-Holland, Amsterdam.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995): Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics* **44**, 455–472.

Gourieroux, C. (2006): Continuous time Wishart process for stochastic risk. *Econometric Reviews* **25**, 177–217.

Gourieroux, C., Jasiak, J. and Sufana, R. (2004): The Wishart autoregressive process of multivariate stochastic volatility. *Discussion paper: University of Toronto*.

Gupta, A. K., González-Farías, G. and Domínguez-Molina, J. A. (2004): A multivariate skew normal distribution. *Journal of Multivariate Analysis* **89**, 181–190.

Han, Y. (2006): The economics value of volatility modelling: Asset allocation with a high dimensional dynamic latent factor multivariate stochastic volatility model. *Review of Financial Studies* **19**, 237–271.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994): Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–264.

Harvey, A. C. and Shephard, N. (1996): Estimation of asymmetric stochastic volatility model for asset returns. *Journal Journal of Business and Economic Statistics* **14**, 429–434.

Hörmann, W., Leydold, J. and Derflinger, G. (2004): *Automatic Nonuniform Random Variate Generation*. Springer, Berlin.

Jacquier, E., Polson, N. G. and Rossi, P. E. (1994): Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business and Economic Statistics* **12**, 371–389.

Jacquier, E., Polson, N. G. and Rossi, P. E. (1999): Stochastic volatility: Univariate and multivariate extensions. *CIRANO Working paper 99s–26, Montreal*.

Jungbacker, B. and Koopman, S. J. (2006): Monte Carlo likelihood estimation for three multivariate stochastic volatility models. *Econometric Reviews* **25**, 385–408.

Kawakatsu, H. (2006): Matrix exponential GARCH. *Journal of Econometrics* **134**, 95–128.

Kim, S., Shephard, N. and Chib, S. (1998): Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**, 361–393.

King, M., Sentana, E. and Wadhwani, S. (1994): Volatility and links between national stock markets. *Econometrica* **62**, 901–933.

Liesenfeld, R. and Richard, J.-F. (2003): Univariate and multivariate stochastic volatility models: Estimation and diagnostics. *Journal of Empirical Finance* **10**, 505–531.

Lopes, H. F. and Carvalho, C. M. (2007): Factor stochastic volatility with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference* **137**, 3082–3091.

Omori, Y., Chib, S., Shephard, N. and Nakajima, J. (2007): Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics* **140**, 425–449.

Philipov, A. and Glickman, M. E. (2006a): Factor multivariate stochastic volatility via Wishart processes. *Econometric Reviews* **25**, 311–334.

Philipov, A. and Glickman, M. E. (2006b): Multivariate stochastic volatility via Wishart processes. *Journal of Business and Economic Statistics* **24**, 313–328.

Pitt, M. K., Chan, D. and Kohn, R. (2006): Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537–554.

Pitt, M. K. and Shephard, N. (1999): Time varying covariances: a factor stochastic volatility approach. *In: Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (Eds.): Bayesian Statistics* **6**, 547–570. Oxford University Press, Oxford.

Protassov, R. S. (2004): EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ. *Statistics and Computing* **14**, 67–77.

Quintana, J. M. and West, M. (1987): An analysis of international exchange rates using multivariate DLMs. *The Statistician* **36**, 275–281.

Ray, B. K. and Tsay, R. S. (2000): Long-range dependence in daily stock volatilities. *Journal of Business and Economic Statistics* **18**, 254–262.

Schmidt, R., Hrycej, T. and Stützle, E. (2006): Multivariate distribution models with generalized hyperbolic margins. *Computational Statistics and Data Analysis* **50**, 2065–20096.

Shephard, N. (2004): *Stochastic Volatility: Selected Readings*. Oxford University Press, Oxford.

Shephard, N. and Pitt, M. K. (1997): Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–667.

Smith, M. and Pitts, A. (2006): Foreign exchange intervention by the Bank of Japan: Bayesian analysis using a bivariate stochastic volatility model. *Econometric Reviews* **25**, 425–451.

So, M. K. P. and Kwok, W. Y. (2006): A multivariate long memory stochastic volatility model. *Physica A* **362**, 450–464.

So, M. K. P., Li, W. K. and Lam, K. (1997): Multivariate modelling of the autoregressive random variance process. *Journal of Time Series Analysis* **18**, 429–446.

So, M. K. P., Lam, K. and Li, W. K. (1998): A stochastic volatility model with Markov switching. *Journal of Business and Economic Statistics* **16**, 244–253.

Tims, B. and Mahieu, R. (2006): A range-based multivariate stochastic volatility model for exchange rates. *Econometric Reviews* **25**, 409–424.

Tsay, R. S. (2005). *Analysis of Financial Time Series: Financial Econometrics* (2nd ed.). Wiley, New York.

Watanabe, T. and Omori, Y. (2004): A multi-move sampler for estimating non-Gaussian times series models: Comments on Shephard and Pitt (1997). *Biometrika* **91**, 246–248.

Wong, F., Carter, C. and Kohn, R. (2003): Efficient estimation of covariance matrix selection models. *Biometrika* **90**, 809–830.

Yu, J. (2005): On leverage in a stochastic volatility model. *Journal of Econometrics* **127**, 165–178.

Yu, J. and Meyer, R. (2006): Multivariate stochastic volatility models: Bayesian estimation and model comparison. *Econometric Reviews* **25**, 361–384.

An Overview of Asset–Price Models

Peter J. Brockwell *

Abstract Discrete-parameter time-series models for financial data have received, and continue to receive, a great deal of attention in the literature. Stochastic volatility models, ARCH and GARCH models and their many generalizations, designed to account for the so-called stylized features of financial time series, have been under development and refinement now for some thirty years. At the same time there has been a rapidly developing interest in continuous-time models, largely as a result of the very successful application of stochastic differential equation models to problems in finance, exemplified by the derivation of the Black-Scholes-Merton (BSM) option-pricing formula and its generalizations. In this overview we start with the BSM option-pricing model in which the asset price is represented by geometric Brownian motion. We then discuss the limitations of the model and survey the various models which have been proposed to provide more realistic representations of empirically observed asset prices. In particular, the observed non-Gaussian distributions of log returns and the appearance of sharp changes in log asset prices which are not consistent with Brownian motion paths have led to an upsurge of interest in Lévy processes and their applications to financial modelling.

Peter J. Brockwell
Department of Statistics, Colorado State University, Fort Collins, Colorado, 80523-1877, U.S.A., e-mail: pjbrock@stat.colostate.edu

* I am indebted to the National Science Foundation for support of this work under the grant DMS-0744058 and to Alexander Lindner for valuable comments on the manuscript.

DOI: 10.1007/978-3-540-71297-8_17,

1 Introduction

For approximately thirty years now, discrete-time models (including stochastic volatility, ARCH, GARCH and their many generalizations) have been developed to reflect the so-called *stylized features* of financial time series. These properties, which include tail heaviness, volatility clustering and serial dependence without correlation, cannot be captured with traditional linear time series models. If S_n denotes the price of a stock or other financial asset at time $n,\ n = 0, 1, 2, \ldots$, then the series of log returns, $\{\log S_n - \log S_{n-1}, n \in \mathbb{N}\}$, is typically represented by either a discrete-time stochastic volatility model or a GARCH process. These models have been studied intensively since their introduction and a variety of parameter estimation techniques have been developed. For an excellent review and comparison of these models see Shephard (1996). For a more recent account of GARCH processes see the article of Lindner (2008) in the current volume, and for stochastic volatility models see the article of Davis and Mikosch (2008). Apart from the need to develop models which capture the distinctive features of financial time series, much of the motivation for developing these models derives from the key role played by volatility in the pricing of options and the need to understand, quantify and forecast its evolution in time.

In mathematical finance, most of the theoretical developments in the pricing of contingent claims (or options) have been made in a continuous-time framework, thanks to the power of Itô calculus, Girsanov's theorem, martingale methods, and other tools associated with the analysis of stochastic differential equations. In fact these developments, which permit the analysis of quite complicated ('exotic') options, have also been a powerful stimulus for the popularization and development of stochastic calculus itself. The celebrated work of Black and Scholes (1973) and Merton (1973) was based on a geometric Brownian motion model for the asset price $S(t)$ at time t (see (1.1) below). Their results, besides winning the Nobel Economics Prize for Merton and Scholes in 1997 (unfortunately Black died before the award was made), inspired an explosion of interest, not only in the pricing of more complicated financial derivatives, but also in the development of new continuous-time models which, like the discrete-time ARCH, GARCH and stochastic volatility models, better reflect the observed properties of financial time series. This development has resulted in a variety of models which are the subject of the articles in this section of the Handbook.

In addition to their central role in option-pricing, time series models with continuous time parameter are particularly well-suited to modelling irregularly spaced data (see e.g. Jones (1985)). Lévy-driven continuous-time autoregressive moving average (CARMA) models play a role in continuous time analogous to that of ARMA models in discrete time, allowing a very flexible range of autocorrelations and marginal distributions, suitable in particular for the modelling of volatility as a continuous-time stationary series (see the article by Brockwell (2008) in this volume).

The use of continuous-time models in finance goes back to Bachelier (1900), who used Brownian motion to represent the prices $\{S(t), t \geq 0\}$ of a stock in the Paris Bourse. This model had the unfortunate feature of permitting negative stock prices, a shortcoming which was eliminated in the geometric Brownian motion model of Samuelson (1965), according to which $S(t)$ satisfies the Itô equation,

$$\mathrm{d}S(t) = \mu S(t)\ \mathrm{d}t + \sigma S(t)\ \mathrm{d}W(t) \text{ with } S(0) > 0. \tag{1.1}$$

In this equation $\{W(t), t \geq 0\}$ is standard Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$ with filtration $\{\mathcal{F}_t\}$ where $\mathcal{F}_t$ is the sub-σ-algebra of $\mathcal{F}$ generated by $\{W(s), 0 \leq s \leq t\}$ and the null sets of $\mathcal{F}$. The solution of (1.1) satisfies

$$S(t) = S(0) \exp\left[(\mu - \sigma^2/2)t + \sigma W(t)\right], \tag{1.2}$$

so that the log asset price in this model is Brownian motion and the log return over the time-interval $(t, t + \Delta)$ is

$$\log \frac{S(t+\Delta)}{S(t)} = (\mu - \frac{1}{2}\sigma^2)\Delta + \sigma(W(t+\Delta) - W(t)).$$

For disjoint intervals of length Δ the log returns are therefore independent normally distributed random variables with mean $(\mu - \sigma^2/2)\Delta$ and variance $\sigma^2\Delta$. The normality of the log returns is a conclusion which can easily be checked against observed returns, and it is found that the deviations are substantial for time intervals of the order of a day or less, becoming less apparent as Δ increases. This is one of the reasons for developing the models described in later sections.

The parameter σ^2 in (1.1) is called the *volatility* parameter and its significance for option pricing was clearly demonstrated in the pricing by Black, Scholes and Merton of a European call option. Such an option, if sold at time 0, gives the buyer the right, but not the obligation, to buy one unit of the stock (with market price satisfying (1.1)) at the *strike time* T for the *strike price* K. At time T the option has the cash value $h(S(t)) = \max(S(T) - K, 0)$ since the option will be exercised only if $S(T) > K$, in which case the holder of the option can buy the stock at the price K and resell it instantly for $S(T)$. However it is not clear at time 0, since $S(T)$ is random, what price the buyer should pay for this privilege. Assuming (i) the existence of a risk-free asset with price process,

$$B(t) = B(0) \exp(rt),\ r > 0, \tag{1.3}$$

(ii) the ability to buy and sell arbitrary (positive or negative) amounts of the stock and the risk-free asset continuously with no transaction costs, and (iii) an arbitrage-free market (i.e., a market in which it is impossible to make a non-negative profit which is strictly positive with probability greater than zero), Black, Scholes and Merton showed that there is a unique *fair price* for

the option in the sense that both higher and lower prices introduce demonstrable arbitrage opportunities. Details of the derivation can be found in most books dealing with mathematical finance (e.g. Campbell, Lo and McKinlay (1996), Mikosch (1998), Steele (2001), Shreve (2004), Björk (2004) and Klebaner (2005)). In the following paragraphs we give a sketch of two arguments, following Mikosch (1998), leading to this fair price for the Black-Scholes-Merton (henceforth BSM) model.

In the first argument, we attempt to construct a self-financing portfolio, consisting at time t of a_t shares of the stock and b_t shares of the risk-free asset, where a_t and b_t are random variables measurable with respect to $\mathcal{F}_t$. We require the value of this portfolio at time t, namely

$$V(t) = a_t S(t) + b_t B(t), \tag{1.4}$$

to satisfy the self-financing condition,

$$\mathrm{d}V(t) = a_t \, \mathrm{d}S(t) + b_t \, \mathrm{d}B(t), \tag{1.5}$$

and to match the value of the option at time T, i.e.,

$$V(T) = h(S(T)) = \max(S(T) - K, 0). \tag{1.6}$$

If such an *investment strategy*, $\{(a_t, b_t), 0 \le t \le T\}$ can be found, then $V(0)$ must be the fair value of the option at the purchase time $t = 0$. A higher price for the option would allow the seller to pocket the difference δ and invest the amount $V(0)$ in such a way as to match the value of the option at time T. Then at time T, if $S(T) < K$ the option will not be exercised and the portfolio and the option will both have value zero. If $S(T) > K$ the seller sells the portfolio for $S(T)-K$, then buys one stock for $S(T)$ and receives K for it from the holder of the option. Since there is no loss involved in this transaction, the seller is left with a net profit of δ. The seller of the option therefore makes a non-negative profit which is strictly positive with non-zero probability, in violation of the no arbitrage assumption. Similarly a lower price than $V(0)$ would create an arbitrage opportunity for the buyer. In order to determine $V(t)$, a_t and b_t we look for a smooth function $v(t, x),\ t \in [0, T],\ x > 0$, such that

$$V(t) = v(t, S(t)),\ t \in [0, T], \tag{1.7}$$

satisfies the conditions (1.5) and (1.6). Equating the expressions for $V(t) - V(0)$ obtained by applying Itô calculus to (1.5) and (1.7), we find that $a_t = \frac{\partial v}{\partial x}(t, S(t))$, where the function v must satisfy the partial differential equation,

$$\frac{\partial v}{\partial t} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 v}{\partial x^2} + r\frac{\partial v}{\partial x} = rv, \tag{1.8}$$

with boundary condition,

$$v(T,x) = h(x) = \max(x - K, 0), \tag{1.9}$$

which, with (1.8), uniquely determines the function v and hence $V(t)$, a_t and $b_t = (V(t) - a_t S(t))/B(t)$ for each $t \in [0,T]$. Thus we have arrived at an investment strategy $\{(a_t, b_t), 0 \le t \le T\}$ which satisfies (1.5) and (1.6) and which, under the assumed idealized trading conditions, can be implemented in practice. Since at time T this portfolio has the same value as the option, $V(0)$ must be the fair value of the option at time $t = 0$; otherwise an arbitrage opportunity would arise. The option is said to be *hedged* by the investment strategy $\{(a_t, b_t)\}$. A key feature of this solution (apparent from (1.8) and (1.9)) is that both the strategy and the fair price of the option are independent of μ, *depending on S only through the volatility parameter σ^2*.

A particularly elegant and powerful way of arriving at the solution of (1.8) and (1.9) is to use a second argument, based on the fact that for the BSM model there is a unique probability measure Q which is equivalent to the original probability measure P (i.e., it has exactly the same null sets) and which, when substituted for P in the probability space on which the stock prices are defined, causes the discounted price process $\{e^{-rt}S(t)\}$ to become a *martingale*, i.e., to satisfy the condition,

$$E_Q(e^{-rt}S(t)|\mathcal{F}_s) = e^{-rs}S(s) \text{ for all } s \le t, \tag{1.10}$$

where E_Q denotes expectation with respect to the new probability measure Q. The probability measure Q is called the *equivalent martingale measure* or EMM. It is unique for the BSM model, but for other models the questions of its existence and uniqueness become serious issues.

The martingale-based argument leading to the BSM pricing formula is as follows. Itô's formula applied to the discounted price process,

$$\tilde{S}(t) := e^{-rt}S(t),$$

gives

$$\frac{\mathrm{d}\tilde{S}(t)}{\tilde{S}(t)} = (\mu - r)\mathrm{d}t + \sigma \mathrm{d}W(t) = \sigma \mathrm{d}\tilde{W}(t), \tag{1.11}$$

where $\tilde{W}(t) := (\mu - r)t/\sigma + W(t)$. The solution of (1.11) satisfies

$$\tilde{S}(t) = \tilde{S}(0)e^{\sigma\tilde{W}(t) - \sigma^2 t/2},$$

which is an $\{\mathcal{F}_t\}$-martingale if $\{\tilde{W}(t), 0 \le t \le T\}$ is standard Brownian motion adapted to $\{\mathcal{F}_t\}$. However by Girsanov's theorem this is the case under the probability measure Q whose Radon-Nikodym derivative with respect to P is

$$\frac{\mathrm{d}Q}{\mathrm{d}P} = \exp\left(-\frac{\mu - r}{\sigma}W(T) - \frac{(\mu - r)^2}{2\sigma^2}T\right). \tag{1.12}$$

Assuming the existence of a portfolio (1.4) which satisfies the self-financing condition (1.5) and the boundary condition (1.6), the discounted portfolio value is

$$\tilde{V}(t) = e^{-rt}V(t). \tag{1.13}$$

Applying Itô's formula to this expression we obtain

$$\mathrm{d}\tilde{V}(t) = -r\tilde{V}(t)\mathrm{d}t + e^{-rt}\mathrm{d}V(t) = a_t\ \mathrm{d}\tilde{S}(t),$$

and hence, from (1.11),

$$\tilde{V}(t) = \tilde{V}(0) + \int_0^t a_s \mathrm{d}\tilde{S}(s) = V(0) + \sigma\int_0^t a_s\tilde{S}(s)\mathrm{d}\tilde{W}(s). \tag{1.14}$$

Since $a_t\tilde{S}(t) \in \mathcal{F}_t$ for each $t \in [0,T]$ and, under the probability measure Q, $\tilde{W}$ is Brownian motion adapted to $\{\mathcal{F}_t\}$, we conclude that $\tilde{V}$ is an $\{\mathcal{F}_t\}$-martingale. Hence

$$\tilde{V}(t) = E_Q[\tilde{V}(T)|\mathcal{F}_t],\ t \in [0,T],$$

and

$$V(t) = e^{rt}\tilde{V}(t) = E_Q[e^{-r(T-t)}h(S(T))|\mathcal{F}_t], \tag{1.15}$$

where $h(S(T))$ is the value of the option at time T. For the European call option $h(S(T)) = \max(S(T)-K,0)$ and a straightforward calculation using (1.15) gives, in the notation of (1.7),

$$v(t,x) = x\Phi(z_1) - Ke^{-r(T-t)}\Phi(z_2), \tag{1.16}$$

where Φ is the standard normal cumulative distribution function,

$$z_1 = \frac{\log(x/K) + (r+\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \text{ and } z_2 = z_1 - \sigma\sqrt{T-t}.$$

The quantity $m = (\mu - r)/\sigma$ which appears in the Radon-Nikodym derivative $\mathrm{d}Q/\mathrm{d}P$ is called the *market price of risk* and represents the excess, in units of σ, of the instantaneous rate of return μ of the risky asset S over that of the risk-free asset B. If $m = 0$ then $Q = P$ and the model is said to be *risk-neutral*.

Although the model (1.1) has many shortcomings as a representation of asset prices, the remarkable achievement of Black, Scholes and Merton in using it to derive a unique arbitrage-free option price has inspired enormous interest and progress in the field of financial mathematics. As a result of their pioneering work, research in continuous-time financial models has blossomed, with much of it directed at the construction, estimation and analysis of more realistic continuous-time models for the evolution of stock prices, and the pricing of options based on such models. In the following sections we summa-

rize the limitations of the BSM model and briefly discuss some of the models which have been developed to provide more realistic representations of the empirical data and to permit the analysis of more complicated contingent claims.

2 Shortcomings of the BSM Model

Under the model (1.1) the sample-paths of the log asset prices are those of Brownian motion. As already indicated, this implies that, for each fixed $\Delta > 0$, the log returns $\{\log S((n+1)\Delta) - \log S(n\Delta), n = 0, 1, 2, \ldots\}$ are independent and identically distributed Gaussian random variables. However inspection of empirical log asset prices, especially at time intervals Δ of one day or less, reveals significant negative skewness of the distribution of these increments and kurtosis which is significantly higher than the value 3, appropriate for normally distributed random variables. In order to reflect these observations we need to consider models for which the marginal distribution of the increments are non-Gaussian.

Moreover, although the observed increments typically exhibit no significant sample correlations, their squares and absolute values usually have autocorrelation functions which are significantly different from zero, indicating the need for models in which the increments are not independent as expected under the BSM model.

The observed increments appear also not to be identically distributed. Their estimated variances change with time in an apparently random manner. Assuming the validity of the BSM model it is possible to estimate the parameter σ^2 per trading day on day n by computing the sum of squares

$$\hat{\sigma}_n^2 := \sum_{i=1}^{N} \left(\log S_n \left(\frac{i}{N} \right) - \log S_n \left(\frac{i-1}{N} \right) \right)^2, \tag{2.1}$$

where the summands are the squared increments of the log price over intervals obtained by breaking the day into intervals of length $1/N$ days with N large. The sequence $\hat{\sigma}_n^2$ is known as the *realized volatility* per day. It is found in practice to vary significantly from one day to the next. The sequence $\{\hat{\sigma}_n^2\}$ of realized volatilities exhibits clustering, i.e., periods of low values interrupted by bursts of large values, and has the appearance of a positively correlated stationary sequence, reinforcing the view that volatility is not constant as in the BSM model and suggesting the need for a model in which volatility is stochastic. Such observations are precisely those which led to the development in discrete time of stochastic volatility, ARCH and GARCH models, and suggest the need for analogous models with continuous time parameter.

If we were to assume the validity of the BSM model for stock prices and also to adopt the widely-held view that in real-world markets there is no arbitrage,

then the quoted price of stock options should be exactly as computed in Section 1. This argument provides us with a consistency check on the BSM model. Given the time to maturity $T-t$, the strike price K and the quoted price $Q(t)$ for a European call option at time t, then from the risk-free interest rate r and the price at time t of the stock, equation (1.16) can be used to calculate the *implied volatility*, $\sigma_t^2(K,T)$ at time t. If the BSM model is appropriate the implied volatility should be independent of K and T, however it is found in practice to depend on both. If we plot $\sigma_t(\cdot,T)$ for fixed t and T the graph usually has the appearance of a smile, the so-called *volatility smile*. The non-constancy of implied volatility is another indicator of the need to improve on the BSM model.

Finally, if we compare daily stock prices with daily values of simulated Brownian motion having the corresponding estimated drift and volatility parameters, we find that the stock prices exhibit occasional jumps which are much larger than the daily increments of the Brownian paths. This suggests that a good model for stock prices and log stock prices should allow for the possibility of jumps.

3 A General Framework for Option Pricing

The martingale argument of Section 1 was extended by Harrison and Pliska (1981), to the much more general model in which the processes $\{S(t)\}$ and $\{B(t)\}$ are semimartingales and the claim at time T, instead of having the form $h(S(T))$, is a non-negative random variable $X \in \mathcal{F}_T$ with $EX < \infty$, where $\{\mathcal{F}_t, 0 \le t \le T\}$ is the filtration generated by $\{(S(t), B(t))\}$ and the P-null sets of $\mathcal{F}$. This allows for path-dependent claim functions such as $\max_{t\in[0,T]} S(t)$. In this more general setting however the existence and uniqueness of an equivalent martingale measure is not always guaranteed and pricing on the basis of no arbitrage is not always possible. In the definition (1.5) of a self-financing strategy $\{(a_t, b_t)\}$ the processes $\{a_t\}$ and $\{b_t\}$ are required to be predictable processes such that the integrated form of (1.5) is well-defined. The self-financing condition is then equivalent to the condition that the discounted price process, $\{V(t)/B(t)\}$, has the representation,

$$\frac{V(t)}{B(t)} = V(0) + \int_0^t a_u dZ(u), \tag{3.1}$$

where $Z(t)$ is the discounted stock price $S(t)/B(t)$.

For the remainder of this section we shall assume that there is (at least one) EMM Q, i.e., a probability measure on $\mathcal{F}$ with the same null-sets as P under which $Z(t)$ is a martingale. Under this assumption the model presents no arbitrage possibility. This result is often called the *first fundamental theorem*

of asset pricing. With a weaker definition of arbitrage than the one we have given, the converse also holds (see Delbaen and Schachermayer (1994)).

Let $\mathcal{L}(Z)$ be the class of predictable processes H such that the process $\{\sqrt{\int_0^t H_u^2 d[Z,Z](u)}\}$ is locally integrable under Q, where $[Z,Z]$ is the quadratic variation process of Z.

An *admissible* strategy is defined to be a predictable self-financing strategy such that $a \in \mathcal{L}(Z)$ and $\{V(t)/B(t)\}$ is a non-negative Q-martingale.

A claim X at time T is said to be *attainable* if there is an admissible strategy $\{(a_t, b_t)\}$ such that $V(T) = X$. This means that there is an admissible strategy which replicates the value of the claim at time T and for which the corresponding discounted price process $V(t)/B(t)$ is a Q-martingale. Consequently, in order to avoid arbitrage, the fair value of the option at time $t < T$ must then be

$$V(t) = B(t)E_Q(X/B(T)|\mathcal{F}_t). \tag{3.2}$$

The following result enables us to identify the attainable claims. If X is an integrable claim (i.e., if $E_Q(X/B(T)) < \infty$), then X is attainable if and only if the process $M(t) := E_Q(X/B(T)|\mathcal{F}_t)$ has the representation,

$$M(t) = M(0) + \int_0^t H(u)dZ(u), \text{ for some } H \in \mathcal{L}(Z). \tag{3.3}$$

The model is said to be *complete* if every integrable claim X is attainable. But this is the same as saying that the discounted price process $Z(t)$ has the predictable representation property, i.e., that *every* martingale has a representation (3.3) for some $H \in \mathcal{L}(Z)$. A necessary and sufficient condition for this is that the equivalent martingale measure Q is unique. This is sometimes called the *second fundamental theorem of asset pricing*.

4 Some Non-Gaussian Models for Asset Prices

The preceding section provides a very general framework for the arbitrage-free pricing of a contingent claim X based on a single stock and a money-market account when there exists a unique equivalent martingale measure Q. The fair price at time zero, when the option is purchased, is, from (3.2),

$$V(0) = B(0)E_Q(X/B(T)). \tag{4.1}$$

The expectation in (4.1) cannot generally be calculated analytically, in spite of the elegant solution for the BSM model, however it does permit the estimation of $V(0)$ by Monte-Carlo simulation once the process $\{(S(t), B(t))\}$ has been specified. In this section we consider some of the models which have been proposed for $\{S(t)\}$, in order to address the limitations of the BSM model listed in Section 2.

The first of these is the diffusion model obtained by replacing the constant parameters μ, σ and r in (1.1) and (1.3) by functions $\mu(t, S(t))$, $\sigma(t, S(t))$ and $r(t, S(t))$. In this case the first argument used in Section 1 leads to the partial differential equation (1.8) and terminal condition (1.9) with μ, σ and r replaced by the corresponding functions of t and x. The fair value of the option is obtained as before from the solution $v(t, x)$ of this partial differential equation, and the weights a_t and b_t of the self-financing portfolio which replicates $h(X(T))$ can be expressed in terms of v and its derivatives. This family of diffusion models for the stock price S allows for a much greater variety of marginal distributions than the Gaussian marginals of the BSM model, however the sample paths of S are still continuous.

In order to account for the occasional sharp changes observed in the sample-paths of asset prices and at the same time to allow for observed log returns which are not Gaussian, a natural step is to replace the exponent in (1.2) by a Lévy process L, i.e., a process with homogeneous independent increments, continuous in probability, with càdlàg sample-paths and $L(0) = 0$. This leads to the so-called *exponential Lévy model*,

$$S(t) = S(0) \exp(L(t)),$$

whose log returns over time intervals of length 1 have the distribution of $L(1)$, which can be any infinitely divisible distribution. The simplest examples of Lévy processes are Brownian motion, which has continuous sample paths, and the Poisson process, which increases only by jumps of size one. In general a Lévy process can be expressed as the sum of a Brownian motion with drift and an independent pure-jump process. Pure jump Lévy processes and exponential Lévy processes are discussed in detail in the article by Eberlein (2008) in this volume, where examples of Lévy processes which have been found especially useful in financial modelling are also given. For more extensive treatments of Lévy processes, see the books of Applebaum (2004), Bertoin (1996), Protter (2004) and Sato (1999). Except in the Brownian motion and Poisson process cases, the exponential Lévy model is incomplete. There are many equivalent martingale measures and there is no unique arbitrage-based (or risk-neutral) price for an option. The problem of choosing an EMM in this situation, computing the corresponding price of a European option and matching it to prices quoted in the market is discussed in Schoutens (2003). A general account of option pricing, covering the general situation in which there is no unique EMM is contained in the article by Kallsen (2008) in this volume.

Lévy processes also play a key role in the stochastic volatility model of Barndorff-Nielsen and Shephard (2001a, 2001b) (henceforth called BNS model) in which the volatility σ^2 is a stationary Ornstein–Uhlenbeck (O-U) process driven by a non-decreasing Lévy process L, i.e.,

$$\sigma^2(t) = \int_{-\infty}^{t} e^{-\lambda(t-y)} dL(\lambda y), \tag{4.2}$$

where $\lambda > 0$. The log asset price $G(t) = \ln P(t)$ satisfies an equation of the form

$$dG_t = (\mu - \frac{1}{2}\sigma^2(t))dt + \sigma(t)\, dW(t) + \rho dL(\lambda t), \tag{4.3}$$

where W is a Brownian motion independent of the Lévy process and ρ is a non-positive real parameter which accounts for the so-called *leverage effect*. The autocorrelation function of the process σ^2 is $\rho(h) = \exp(-\lambda|h|), h \in \mathbb{R}$, with $\lambda > 0$, but this class of functions can be extended by specifying the volatility to be a superposition of O-U processes as in Barndorff–Nielsen (2001), or a Lévy-driven CARMA (continuous-time ARMA) process as in Brockwell (2004). The BNS model defined by (4.2) and (4.3) is incomplete. There is a family of equivalent martingale measures, the structure of which was studied by Nicolato and Venardos (2003) who argue that it is sufficient to consider the subset of EMM's under which the log returns continue to be described by a BNS model. For such an EMM Q they show how to compute $E_Q(e^{-r(T-t)}h(X(T))|\mathcal{F}_t)$ for a European contract with claim $h(X(T))$. Using gamma and inverse gamma Ornstein-Uhlenbeck processes and estimating parameters by minimizing the mean-squared error between model and market option prices they find that both models perform well when applied to European call options on the S&P500 index, giving good matches between the observed and fitted volatility smiles. The book of Schoutens (2003) discusses option pricing also for a class of stochastic volatility models in which the stock price is the exponential of a stochastically time-changed Lévy process. Simulation methods and the pricing of exotic options are also discussed.

In view of the wide use of discrete-time ARCH and GARCH models for asset prices, a great deal of research has been devoted to the development of analogous continuous-time models. An early attempt to bridge the gap between discrete-time GARCH models and continuous-time models resulted in the GARCH(1,1) diffusion approximation of Nelson (1990). An outline of the argument used by Nelson is given by Lindner (2008) in this volume. See also Drost and Werker (1996) and Duan (1997). As in the continuous-time stochastic volatility models we model the logarithm of the asset price itself, i.e. $G(t) = \log S(t)$, rather than its increments as in discrete time. Nelson's diffusion limit for the log asset price and squared volatility is the unique solution $\{(G(t), \sigma^2(t)),\ t \geq 0\}$ of the equations,

$$\mathrm{d}G(t) = \sigma(t)\, \mathrm{d}W^{(1)}(t), \quad \mathrm{d}\sigma^2(t) = \theta(\gamma - \sigma^2(t)) + \rho\sigma^2(t)\, \mathrm{d}W^{(2)}(t), \tag{4.4}$$

with initial value $(G(0), \sigma^2(0))$, where $W^{(1)}$ and $W^{(2)}$ are independent standard Brownian motions and ω, λ and θ are parameters (see Lindner (2008) for details). This model for G differs fundamentally from the GARCH(1,1) model in that it is driven by two independent processes instead of one and the squared volatility evolves independently of $W^{(1)}$. The behaviour of this diffusion limit is therefore rather different from that of a GARCH process (see Lindner (2008)).

A different approach to constructing a continuous-time analogue of the GARCH(1,1), the COGARCH(1,1) process, was taken by Klüppelberg et al. (2004). The starting point was the explicit expression for the volatility of the discrete-time GARCH(1,1) process which can be computed recursively from the difference equations,

$$\sigma_n^2 = \alpha_0 + \beta_1 \sigma_{n-1}^2 + \alpha_1 e_{n-1}^2 \sigma_{n-1}^2, \tag{4.5}$$

where $\alpha_0 > 0$, $\alpha_1, \beta_1 \geq 0$, $\alpha_1 + \beta_1 \leq 1$ and $\{e_t,\ t = 1, 2, \ldots\}$ is an iid sequence with mean 0 and variance 1. This expression is written as an integral and the noise sequence replaced by the jumps of a Lévy process. Details of the construction are contained in Lindner (2008). For GARCH(p, q) processes of higher order, there is no analogue of the explicit expression for σ_n^2, however the process $\{\sigma_n^2\}$ can be regarded as a "self-exciting" ARMA$(q-1, p)$ process driven by the sequence $\{e_{n-1}^2 \sigma_{n-1}^2\}$. This can be clearly seen in equation (4.5) where $p = q = 1$. The COGARCH(p, q) process (with $p \leq q$) is obtained by replacing the self-exciting ARMA$(q, p-1)$ equation for σ_n^2 by a corresponding self-exciting continuous-time ARMA$(q, p-1)$ equation driven by a continuous time analogue of the sequence $\{e_{n-1}^2 \sigma_{n-1}^2\}$. Details can be found in Brockwell et al. (2006) and Brockwell (2008). COGARCH processes with a stationary volatility process have properties that are closely analogous to those of discrete-time GARCH processes. In particular if $G_t^{(r)}$ denotes the increment $G(t+r) - G(t)$ then, under conditions ensuring the finiteness of $E[G_t^{(r)}]^4$, $G_t^{(r)}$ has zero mean, $\{G_{t+h}^{(1)}), h = 0, 1, 2, \ldots\}$ is an uncorrelated sequence and the corresponding sequence of squared increments has the autocovariance function of an ARMA process, while the process $\{\sigma_t^2\}$ has the autocovariance function of a continuous-time ARMA process.

The COGARCH(1,1) process with stationary volatility has been shown to have many of the features of the discrete time GARCH(1,1) process. As shown in Klüppelberg et al. (2004, 2006), the COGARCH(1,1) process has uncorrelated increments, while the autocorrelation functions of the volatility σ^2 and of the squared increments of G decay exponentially. Further, the COGARCH(1,1) process has heavy tails and volatility clusters at high levels, see Klüppelberg et al. (2006) and Fasen et al. (2005). Cluster behaviour can also be achieved in the stochastic volatility model of Barndorff-Nielsen and Shephard if the driving Lévy process has regularly varying tails. For an overview of extremes of stochastic volatility models, see Fasen et al. (2005). The discrete-time EGARCH model of Nelson (1990) was introduced in order to account for the observation that negative shocks have a greater effect on volatility than positive ones. A continuous-time analogue of the EGARCH model is the ECOGARCH model of Haug and Czado (2007).

A unifying and large family of processes which includes several of those introduced in this section is the family of generalized Ornstein-Uhlenbeck (GOU) processes (Lindner and Maller (2005), Maller, Müller and Szimayer (2008)). A GOU process X is defined, in terms of a bivariate Lévy process

(ξ, η) by

$$X_t = m(1 - e^{-\xi_t}) + e^{-\xi_t} \int_0^t e^{\xi_{s-}} d\eta_s + X_0 e^{-\xi_s}, t \geq 0, \tag{4.6}$$

where X_0 is independent of $\{(\xi_t, \eta_t), t \geq 0\}$. Among the processes in this family are the stochastic volatility model of Barndorff-Nielsen and Shephard, the COGARCH(1,1) process and the GARCH(1,1) limiting diffusion of Nelson. For some of the applications of this family in option pricing, insurance and risk theory see Maller et al. (2008). The extremal behaviour of stationary GOU processes is treated in this volume by Fasen (2008).

5 Further Models

The asset-price models considered in the preceding sections constitute a small but important part of the multitude of continuous-time stochastic models currently of importance in mathematical finance. In this final section we highlight a few of the important classes of models and problems, the details of which cannot be included in this brief overview.

In order to account for dependence between the price processes of different assets, multiple-asset models are required. Shreve (2004) considers option pricing based on the model,

$$dS_i(t) = \alpha_i(t) S_i(t) \mathrm{d}t + S_i(t) \sum_{j=1}^{d} \sigma_{ij}(t) \mathrm{d}W_j(t), \ i = 1, \ldots, m,$$

where the vector $[\alpha_i]_{i=1,\ldots,m}$ and the volatility matrix $[\sigma_{ij}]_{i=1,\ldots,m;j=1,\ldots,d}$ are adapted processes and W_j, $j = 1, \ldots d$, are independent standard Brownian motions. Multivariate generalizations of the BNS model and of the COGARCH(p, q) model have also been developed by Stelzer (2007) and Pigorsch and Stelzer (2007) respectively.

Another large class of models for which there is an extensive literature are those for bonds and interest rates. For an extensive treatment of these see the book of Björk (2004) and the article by Björk (2008) in this volume. For a Lévy based approach see also Eberlein and Raible (1999).

The estimation of volatility itself from high frequency data presents many challenging problems. In equation (2.1) we introduced the notion of realized volatility and, in the context of the BSM model, this converges as $N \to \infty$ to the parameter σ^2 (per day). However in practice, factors such as within-day variation, discreteness of the price structure and the presence of jumps, complicate the choice of N and the interpretation of the realized volatility as defined by (2.1). An extensive discussion of realized volatility is contained in the article of Andersen and Benzoni (2008) in this volume. Realized volatil-

ity series are generally found to exhibit very slowly decaying autocorrelation functions, suggesting the use of long-memory models or continuous-time ARMA models with an autoregressive root close to zero in order to represent them (see, e.g. Todorov (2007)).

The general problem of parameter estimation for continuous-time models is complicated by the fact that observations are always made at discrete times. When the continuous-time process is Markovian and the transition probabilities can be computed it is possible to write down the likelihood of the observations and hence to carry out estimation by maximum likelihood. Except in very special cases however the transition probabilities have no simple explicit form and approximation of the likelihood or alternative methods must be used. The papers of Aït-Sahalia and Mykland (2008) and Phillips and Yu (2008) in this volume address these problems. See also the paper of Kelly et al.(2004).

Estimation for the BNS stochastic volatility model has been carried out by Roberts et al. (2004) and Gander and Stephens (2007) using Markov chain Monte-Carlo methods and estimation for COGARCH(1,1) models by Haug and Czado (2007) using method of moments estimation.

There still remain many intriguing and challenging problems for the modelling of asset prices. The models described in this overview have provided a great deal of insight into the dynamics of price movements and the critical role of market volatility. They have also been of practical value in the pricing of options. Much remains to be discovered however, particularly with regard to the intra-day price movements and the factors affecting them. For the analysis of tick by tick (or ultra-high-frequency) data it is necessary to take into account both the discrete times at which transactions occur and the price changes at each transaction. The autoregressive conditional duration (ACD) model of Engle and Russell (1998) was constructed for this purpose. The analysis of high frequency data casts light on the trading mechanism and the detailed operation (or microstructure) of the market and remains a particularly active area of research. The book of Tsay (2005) contains a clear account, with applications, of such models.

References

Aït-Sahalia, Y. and Mykland, P.A. (2008): Estimating Volatility in the Presence of Microstructure Noise: A Review of the Theory and Practical Considerations. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 576–598. Springer, New York.

Andersen, T. and Benzoni, L. (2008): Realized volatility. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 554–575. Springer, New York.

Applebaum, D. (2004): Lévy Processes and Stochastic Calculus. Cambridge University Press, Cambridge.

Bachelier, L. (1900): Théorie de la spéculation. *Annales de lÉcole Normale Supérieure* **17**, 21–86.

Barndorff-Nielsen, O.E. (2001): Superposition of Ornstein-Uhlenbeck type processes. *Theory Probab. Appl.* **45**, 175–194.

Barndorff-Nielsen, O.E. and Shephard, N. (2001a): Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics (with discussion). *J. Roy. Statist. Soc. Ser. B* **63**, 167–241.

Barndorff-Nielsen, O.E. and Shephard, N. (2001b): Modelling by Lévy processes for financial econometrics. *In: Barndorff-Nielsen, O.E., Mikosch T. and Resnick S. (Eds.): Lévy Processes - Theory and Applications*, 283–318. Birkhäuser, Boston.

Bertoin, J. (1996): *Lévy Processes*. Cambridge University Press, Cambridge.

Björk, Thomas (2004): *Arbitrage Theory in Continuous Time*. Oxford University Press, Oxford.

Björk, Thomas (2008): An overview of interest rate theory. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 614–651. Springer, New York.

Black, F. and Scholes, M. (1973): The pricing of options and corporate liabilities. *J. Political Economy* **81**, 637–654.

Brockwell, P.J. (2004): Representations of continuous-time ARMA processes. *J. Appl. Probab.* **41A**, 375–382.

Brockwell, P.J. (2008): Lévy-driven continuous-time ARMA processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 456–480. Springer, New York.

Brockwell, P.J., Chadraa, E., and Lindner, A. (2006): Continuous-time GARCH processes. *Ann. Appl. Prob.* **16**, 790–826.

Campbell, J., Lo, A. and McKinlay, C. (1996): The Econometrics of Financial Markets. Princeton University Press, Princeton.

Davis, R.A. and Mikosch, T. (2008): Probabilistic properties of stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 255–267. Springer, New York.

Delbaen, F. and Schachermayer, W. (1994): A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* **300**, 463–520.

Drost, F.C. and Werker, B.J.M. (1996): Closing the GARCH gap: Continuous time GARCH modeling. *Journal of Econometrics* **74**, 31–57.

Duan, J.-C. (1997): Augmented GARCH(p,q) process and its diffusion limit. *Journal of Econometrics* **79**, 97–127.

Eberlein, E. (2008): Jump-type Lévy processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 438–455. Springer, New York.

Eberlein, E. and Raible, S. (1999): Term structure models driven by general Lévy processes. *Mathematical Finance* **9**, 31–53.

Engle, R.F. and Russell, J.R. (1998): Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–1162.

Fasen, V. (2008): Extremes of continuous-time processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 652–667. Springer, New York.

Fasen, V., Klüppelberg, C. and Lindner, A. (2005): Extremal behavior of stochastic volatility models. *In: Shiryaev, A., Grossihno, M.A., Oliviera, P.E. and Esquivel, M.L. (Eds.): Stochastic Finance*. Springer, Heidelberg.

Gander, M.P.S. and Stephens, D.A. (2007): Simulation and inference for stochastic volatility models driven by Lévy processes. *Biometrika* **94**, 627–646.

Harrison, J.M. and Pliska, S.R. (1981): Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes Appl.* **11**, 215–260.

Haug, S. and Czado, C. (2007): An exponential continuous-time GARCH process. *J. Appl. Prob.* **44**, 960–976.

Haug, S., Klüppelberg, C., Lindner, A. and Zapp, M. (2007): Method of moment estimation in the COGARCH(1,1) model. *The Econometrics Journal* **10**, 320–341.

Jones, R.H. (1985): Time series analysis with unequally spaced data. *In: Hannan E.J., Krishnaiah P.R. and Rao M.M. (Eds.): Time Series in the Time Domain, Handbook of Statistics* **5**, 157–178. North Holland, Amsterdam.

Kallsen, J. (2008): Option pricing. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 599–613. Springer, New York.

Kelly, L., Platen, E. and Sørensen, M. (2004): Estimation for discretely observed diffusions using transform functions. *J. Appl. Prob.* **41A**, 99–118.

Klebaner, F. (2005): *Introduction to Stochastic Calculus with Applications*. Imperial College Press, London.

Klüppelberg, C., Lindner, A. and Maller, R. (2004): A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour. *J. Appl. Probab.* **41**, 601–622.

Klüppelberg, C., Lindner, A. and Maller, R.A. (2006): Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models. *In: Kabanov Y., Lipster R. and Stoyanov J. (Eds.): From Stochastic Calculus to Mathematical Finance: The Shiryaev Festschrift*, 393–420. Springer, Heidelberg.

Lindner, A. (2008): Continuous-time GARCH processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 481–496. Springer, New York.

Lindner, A. and Maller, R.A. (2005): Lévy integrals and the stationarity of generalised Ornstein-Uhlenbeck processes. *Stoch. Proc. Appl.* **115**, 1701–1722.

Maller, R.A., Müller, G. and Szimayer A. (2008): Ornstein-Uhlenbeck processes and extensions. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 420–437. Springer, New York.

Merton, R. (1973): The theory of rational option pricing. *Bell J. Economics and Management Science* **4**, 141–183.

Mikosch, T. (1998): *Elementary Stochastic Calculus with Finance in View*. World Scientific, Singapore.

Nelson, D. (1990): ARCH models as diffusion approximations. *J. Econometrics* **45**, 7–38.

Nicolato E. and Vernados, E. (2003): Option pricing in stochastic volatility models of Ornstein-Uhlenbeck type. *Mathematical Finance* **13**, 445–466.

Phillips, P.C.B. and Yu, J. (2008): Maximum likelihood and Gaussian estimation of continuous time models in finance. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 497–530. Springer, New York.

Pigorsch, C. and Stelzer, R. (2007): A Multivariate Generalization of the Ornstein-Uhlenbeck Stochastic Volatility Model. *http://www-m4.ma.tum.de/Papers/Stelzer/PigorschStelzerMultOU.pdf*

Protter, P.E. (2004): *Stochastic Integration and Differential Equations*, 2nd edition. Springer, New York.

Roberts, G., Papaspiliopoulos, O. and Dellaportas, P. (2004): Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *J. R. Statist.Soc.B* **66**, 369–393.

Samuelson, P.A. (1965): Rational theory of warrant pricing. *Ind. Management Review* **6**, 13–31.

Sato, K. (1999): Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge.

Schoutens, W. (2003): *Lévy Processes in Finance*. John Wiley and Sons, Chichester.

Shephard, N. (1996): Statistical aspects of ARCH and stochastic volatility. *In: Cox D.R., Hinkley D.V. and Barndorff-Nielsen O.E. (Eds.): Time Series Models in Econometrics, Finance and Other Fields*, 1–67. Chapman and Hall, London.

Shephard, N. and Andersen, T.G. (2008): Stochastic volatility: Origins and Overview. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 233–254. Springer, New York.

Shreve, S.E. (2004): *Stochastic Calculus for Finance II: Continuous-time Models*. Springer, New York.
Steele, J.M. (2001): *Stochastic Calculus and Financial Applications*. Springer, New York.
Stelzer, R. (2007): Multivariate continuous time Lévy-driven GARCH processes. *Preprint*.
Todorov, V. (2007): Econometric analysis of jump-driven stochastic volatility models. *www.kellogg.northwestern.edu/faculty/todorov/htm/papers/jdsv.pdf*
Tsay, R.S. (2005): *Analysis of Financial Time Series*. John Wiley and Sons, Hoboken.

Ornstein–Uhlenbeck Processes and Extensions

Ross A. Maller, Gernot Müller and Alex Szimayer

Abstract This paper surveys a class of Generalised Ornstein-Uhlenbeck (GOU) processes associated with Lévy processes, which has been recently much analysed in view of its applications in the financial modelling area, among others. We motivate the Lévy GOU by reviewing the framework already well understood for the "ordinary" (Gaussian) Ornstein-Uhlenbeck process, driven by Brownian motion; thus, defining it in terms of a stochastic differential equation (SDE), as the solution of this SDE, or as a time changed Brownian motion. Each of these approaches has an analogue for the GOU. Only the second approach, where the process is defined in terms of a stochastic integral, has been at all closely studied, and we take this as our definition of the GOU (see Eq. (12) below).

The stationarity of the GOU, thus defined, is related to the convergence of a class of "Lévy integrals", which we also briefly review. The statistical properties of processes related to or derived from the GOU are also currently of great interest, and we mention some of the research in this area. In practise, we can only observe a discrete sample over a finite time interval, and we devote some attention to the associated issues, touching briefly on such topics as an autoregressive representation connected with a discretely sampled GOU, discrete-time perpetuities, self-decomposability, self-similarity, and the Lamperti transform.

Some new statistical methodology, derived from a discrete approximation procedure, is applied to a set of financial data, to illustrate the possibilities.

Ross A. Maller
School of Finance & Applied Statistics, and Centre for Mathematics & its Applications, Australian National University ACT 0200, Australia, e-mail: Ross.Maller@anu.edu.au

Gernot Müller
Zentrum Mathematik, Technische Universität München, Boltzmannstrasse 3, D-85747 Garching, Germany, e-mail: mueller@ma.tum.de

Alex Szimayer
Fraunhofer-Institut für Techno-und Wirtschaftsmathematik, Fraunhofer-Platz 1, D-67663 Kaiserslautern, Germany, e-mail: alexander.szimayer@itwm.fraunhofer.de

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_18,

1 Introduction

The Ornstein-Uhlenbeck (throughout: OU) process was proposed by Uhlenbeck and Ornstein (1930) in a physical modelling context, as an alternative to Brownian Motion, where some kind of mean reverting tendency is called for in order to adequately describe the situation being modelled. Since the original paper appeared, the model has been used in a wide variety of applications areas. In Finance, it is best known in connection with the Vasicek (1977) interest rate model. References to this (huge) literature are readily available via library and web searches, and we will not attempt to review it all here. However, to set the scene we will briefly discuss the standard (Gaussian) OU process, driven by Brownian Motion, and concentrate thereafter on some extensions that have recently attracted attention, especially in the financial modelling literature.

2 OU Process Driven by Brownian Motion

The (one-dimensional) Gaussian OU process $X = (X_t)_{t\geq 0}$ can be defined as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}X_t = \gamma(m - X_t)\mathrm{d}t + \sigma \mathrm{d}B_t,\ t > 0, \tag{1}$$

where γ, m, and $\sigma \geq 0$ are real constants, and B_t is a standard Brownian Motion (SBM) on $\mathbb{R}$. X_0, the initial value of X, is a given random variable (possibly, a constant), taken to be independent of $B = (B_t)_{t\geq 0}$. The parameter m can be formally eliminated from (1) by considering $X_t^{(m)} := X_t - m$ rather than X, but we will keep it explicit in view of some later applications.

Alternatively, we could define X in terms of a stochastic integral:

$$X_t = m\left(1 - e^{-\gamma t}\right) + \sigma e^{-\gamma t}\int_0^t e^{\gamma s}\mathrm{d}B_s + X_0 e^{-\gamma t},\ t \geq 0. \tag{2}$$

It is easily verified that X as defined by (2) satisfies (1) for any γ, m, σ, and choice of X_0; it is the *unique, strong Markov solution* to (1), cf. Protter (2005, p. 297). The stochastic integral in (2) is well defined and satisfies the properties outlined in Protter (2005), for example. In particular, $M_\cdot := \int_0^\cdot e^{\gamma s}\mathrm{d}B_s$ is a zero-mean martingale with respect to the natural filtration of B, whose quadratic variation is $[M, M]_\cdot = \int_0^\cdot e^{2\gamma s}\mathrm{d}s$. So $M_t = W_{[M,M]_t}$, $t \geq 0$, where W is an SBM (Protter 2005, p. 88). This leads to a third representation for X as a time changed Brownian motion:

$$X_t = m\left(1 - e^{-\gamma t}\right) + \sigma e^{-\gamma t} W_{(e^{2\gamma t}-1)/2\gamma} + X_0 e^{-\gamma t},\ t \geq 0. \tag{3}$$

Basic properties of X are easily derived from (1)–(3). In particular, conditional on X_0, and assuming X_0 has finite variance, X_t is Gaussian with expectation and covariance functions given by

$$EX_t = m\left(1 - e^{-\gamma t}\right) + e^{-\gamma t} EX_0,\ t \geq 0, \tag{4}$$

and

$$\mathrm{Cov}(X_u, X_t) = \frac{\sigma^2}{2\gamma} e^{-\gamma u}\left(e^{\gamma t} - e^{-\gamma t}\right) + e^{-\gamma(u+t)}\mathrm{Var}X_0,\ u \geq t \geq 0. \tag{5}$$

When, and only when, $\gamma > 0$, the limit $\lim_{t\to\infty} \int_0^t e^{-\gamma s}\mathrm{d}B_s$ exists almost surely (a.s.) as a finite random variable, which we can denote as $\int_0^\infty e^{-\gamma s}\mathrm{d}B_s$. Using time reversal (for a fixed $t > 0$, $(B_s)_{0\leq s\leq t}$ has the same distribution as $(B_{t-s})_{0\leq s\leq t}$) we see from (2) that, for each $t \geq 0$, X_t has the same distribution as

$$\widetilde{X}_t := m\left(1 - e^{-\gamma t}\right) + \sigma \int_0^t e^{-\gamma s}\mathrm{d}B_s + X_0 e^{-\gamma t}, \tag{6}$$

so $\lim_{t\to\infty} \widetilde{X}_t$ exists a.s., is finite, and equals $\widetilde{X}_\infty := m + \sigma \int_0^\infty e^{-\gamma s}\mathrm{d}B_s$, when $\gamma > 0$. If X_t is "started with" initial value X_0, having the distribution of $\widetilde{X}_\infty$, and independent of $(X_t)_{t>0}$, then it is strictly stationary in the sense that the random vectors $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ and $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$ have the same distribution, for any $k = 1, 2, \ldots$, $h > 0$, and $0 < t_1 < t_2 < \ldots < t_k < \infty$. In this case we can extend B_t to $(-\infty, 0)$, note that $\widetilde{X}_\infty \stackrel{D}{:=} m + \sigma \int_{-\infty}^0 e^{\gamma s}\mathrm{d}B_s$, independent of $(X_t)_{t\geq 0}$, and take $X_0 := m + \sigma \int_{-\infty}^0 e^{\gamma s}\mathrm{d}B_s$. From (2), then, we can write

$$X_t = m + \sigma e^{-\gamma t} \int_{-\infty}^t e^{\gamma s}\mathrm{d}B_s,\ t \geq 0. \tag{7}$$

Since B_t has stationary independent increments, from (7) we see that X is a stationary, Markovian, Gaussian process, which is continuous in probability. Conversely, any such process is a (stationary) version of a Gaussian OU process.

The "mean reversion" of X to the constant level m when $\gamma > 0$ can be inferred from (1); if X has diffused above m at some time, then the coefficient of the "$\mathrm{d}t$" drift term is negative, so X will tend to move downwards immediately after, with the reverse holding if X is below m at some time.

Definitions (1) and (2) still make sense when $\gamma \leq 0$. When $\gamma = 0$, X reduces to a zero mean Brownian motion (note that the parameter m is unidentified when $\gamma = 0$) and when $\gamma \leq 0$ of course X is not stationary, in fact $|\int_0^t e^{-\gamma s}\mathrm{d}B_s|$ tends to infinity in probability as $t \to \infty$, so this is an "explosive" case.

3 Generalised OU Processes

There are many ways of generalising the Gaussian OU process, but we will concentrate here on a class of generalisations which has particular application in financial modelling, and has been recently studied intensely from this point of view. There are certainly applications of this class in other areas too.

The idea is to replace the $\mathrm{d}t$ and $\mathrm{d}B_t$ differentials in (1) with the differentials of other semimartingales, or, alternatively, replace the exponential function and Brownian motion in (2) or (3) with other processes. These are quite sweeping generalisations, and to keep the analysis manageable we restrict ourselves to a Lévy generalisation. This is already a profound one, and, apart from greatly increasing applicability, introduces many interesting and important analytical considerations, not least to do with the intricacies of the stochastic calculus. We passed over this aspect in Section 2 because the integrals involve only the *continuous* semimartingale B_t, and are relatively easy to handle. A general Lévy process has a *jump* component which requires special attention in the analysis. But the jumps introduce a modelling feature we wish to incorporate since they prove useful in some financial modelling situations, see, e.g., Geman, Madan and Yor (2000). Another aspect that becomes more interesting (and more difficult!) for jump processes is the statistical analysis; we discuss this below.

Before proceeding, we need to recall some properties of Lévy processes.

3.1 Background on bivariate Lévy processes

We refer to Bertoin (1996) and Sato (1999) for basic results and representations concerning Lévy processes (see also Protter 2005, Ch. I, Sect. 4). Univariate Lévy processes are also considered in Brockwell (2008). For the Generalised OU Process (GOU) we need some specialised material on bivariate Lévy processes, which we briefly review now.

The setup is as follows. Defined on $(\Omega, \mathcal{F}, P)$, a complete probability space, a bivariate Lévy process $(\xi_t, \eta_t)_{t\geq 0}$ is a stochastic process in $\mathbb{R}^2$, with càdlàg paths and stationary independent increments, which is continuous in probability. We take $(\xi_0, \eta_0) = (0, 0)$ and associate with (ξ, η) its natural filtration $(\mathcal{F}_t)_{t\geq 0}$, the smallest right-continuous filtration for which $(\xi_t, \eta_t)_{t\geq 0}$ is adapted, completed to contain all P-null sets.

Especially important is the *Lévy exponent*, $\psi(\theta)$, which is defined in terms of the characteristic function of (ξ_t, η_t) via

$$Ee^{\mathrm{i}\langle(\xi_t,\eta_t),\theta\rangle} =: e^{t\psi(\theta)},$$

where $\langle\cdot,\cdot\rangle$ denotes inner product in $\mathbb{R}^2$. For a bivariate Lévy process the exponent is given by the Lévy-Khintchine representation:

$$\psi(\theta) = \mathrm{i}\langle A,\theta\rangle - \tfrac{1}{2}\langle\theta,\Sigma\theta\rangle + \iint_{|(x,y)|\le 1} \left(e^{\mathrm{i}\langle(x,y),\theta\rangle} - 1 - \mathrm{i}\langle(x,y),\theta\rangle\right) \Pi_{\xi,\eta}(\mathrm{d}x,\mathrm{d}y) + \iint_{|(x,y)|>1} \left(e^{\mathrm{i}\langle(x,y),\theta\rangle} - 1\right) \Pi_{\xi,\eta}(\mathrm{d}x,\mathrm{d}y), \quad \text{for } \theta\in\mathbb{R}^2. \tag{8}$$

Here $|\cdot|$ is Euclidian distance in $\mathbb{R}^2$, $A = (A_1, A_2)$ is a nonstochastic 2–vector, $\Sigma = (\sigma_{rs})$ is a nonstochastic 2×2 non-negative definite matrix, and the Lévy measure, $\Pi_{\xi,\eta}$, is a measure on the Borel subsets of $\mathbb{R}^2\setminus\{0\}$, with $\int(|(x,y)|^2\wedge 1)\Pi_{\xi,\eta}(\mathrm{d}x,\mathrm{d}y) < \infty$. Though its value at $(0,0)$ is not relevant, for definiteness, we can take $\Pi_{\xi,\eta}\{(0,0)\} = 0$. In the literature, Lévy processes such that the Lévy measure of any neighbourhood in $\mathbb{R}^2\setminus\{0\}$ whose closure contains 0 is infinite, are often described as having "infinite activity". Such processes have infinitely many jumps in every nonempty time interval, a.s. The remaining Lévy processes, that is, Lévy processes with "finite activity", are compound Poisson processes (possibly with a drift).

The component processes ξ_t and η_t are Lévy processes in their own right, having canonical triplets $(A_\xi, \sigma_{11}, \Pi_\xi)$ and $(A_\eta, \sigma_{22}, \Pi_\eta)$, say, where the Lévy measures are given by

$$\Pi_\xi\{\Lambda\} := \int_{\mathbb{R}} \Pi_{\xi,\eta}\{\Lambda, dy\} \quad \text{and} \quad \Pi_\eta\{\Lambda\} := \int_{\mathbb{R}} \Pi_{\xi,\eta}\{dx,\Lambda\}, \tag{9}$$

for Λ a Borel subset of $\mathbb{R}\setminus\{0\}$, and the centering constants are given by

$$A_\xi := A_1 + \int_{|x|\le 1} x \int_{|y|\ge\sqrt{1-x^2}} \Pi_{\xi,\eta}\{\mathrm{d}x,\mathrm{d}y\},$$

and similarly for A_η.

A (càdlàg) Lévy process has countably many jumps at most, a.s. We set $(\xi_{s-},\eta_{s-}) := \lim_{u\uparrow s}(\xi_u,\eta_u)$ for $s>0$, and denote the jump process by

$$\Delta(\xi,\,\eta)_t := (\Delta\xi_t,\,\Delta\eta_t) = (\xi_t - \xi_{t-},\,\eta_t - \eta_{t-}), \quad t\ge 0$$

(with $(\xi_{0-},\eta_{0-}) = 0$). If Λ is a Borel subset of $\mathbb{R}^2\setminus\{0\}$, then the expected number of jumps of (ξ,η) of (vector) magnitude in Λ occuring during any unit time interval equals $\Pi\{\Lambda\}$, i.e., for any $t>0$,

$$\Pi\{\Lambda\} = E\sum_{t<s\le t+1} 1_{\{(\Delta\xi_s,\Delta\eta_s)\in\Lambda\}}. \tag{10}$$

Corresponding exactly to the decomposition in (8) is the Lévy-Itô representation of the process as a shift vector plus Brownian plus "small jump" plus "large jump" components:

$$(\xi_t,\eta_t) = (A_1,A_2)t + (B_{\xi,t},B_{\eta,t}) + (\xi_t^{(1)},\eta_t^{(1)}) + (\xi_t^{(2)},\eta_t^{(2)}). \tag{11}$$

Here $(B_{\xi,t}, B_{\eta,t})_{t\geq 0}$ is a Brownian motion on $\mathbb{R}^2$ with mean $(0,0)$ and covariance matrix $t\Sigma$, $(\xi_t^{(1)}, \eta_t^{(1)})_{t\geq 0}$ is a discontinuous (pure jump) process with jumps of magnitude not exceeding 1, which may be of bounded variation on compact time intervals (that is, $\sum_{0<s\leq t} |\Delta(\xi,\eta)_s| < \infty$ a.s. for all $t > 0$), or of unbounded variation, and $(\xi_t^{(2)}, \eta_t^{(2)})_{t\geq 0}$ is a pure jump process with jumps of magnitude always exceeding 1; thus it is a compound Poisson process. The truncation point "1" is arbitrary and can be replaced by any other positive number at the expense only of redefining the shift vector (A_1, A_2). The representation (11) is a great aid to intuition as well as being indispensible in many analyses.

The couple $(\xi_t^{(1)}, \eta_t^{(1)})$ (also a bivariate Lévy process, as is $(\xi_t^{(2)}, \eta_t^{(2)})$) have finite moments of all orders, and by adjusting the centering vector A if necessary we can take $E\xi_1^{(1)} = E\eta_1^{(1)} = 0$. Moments of $\xi_t^{(2)}$ and $\eta_t^{(2)}$ are not necessarily finite; conditions for this to be so, in terms of the canonical measures, are in Sato (1999, p.159, and p.163, ff.). In general, one or more of the components on the righthand side of (11) may not be present, i.e., degenerates to 0. The bivariate Lévy then correspondingly degenerates to a simpler form.

3.2 Lévy OU processes

As a starting point for the generalisation we could use (1), (2), or (3). In our general setting these three definitions do not produce the same process. Each is interesting in its own right, but what is presently known in the literature as the *Generalised OU process* proceeds from (2), and we will adhere to this usage. Thus, we take a bivariate Lévy process (ξ, η) and write

$$X_t = m\left(1 - e^{-\xi_t}\right) + e^{-\xi_t} \int_0^t e^{\xi_{s-}} \mathrm{d}\eta_s + X_0 e^{-\xi_t},\ t \geq 0, \tag{12}$$

where X_0 is independent of $(\xi_t, \eta_t)_{t\geq 0}$, and assumed $\mathcal{F}_0$-measurable. Considerations of the stochastic calculus require us to be precise in specifying the filtration with respect to which the integral in (12) is defined, and we take it to be the natural filtration $(\mathcal{F}_t)_{t\geq 0}$. The Lévy processes ξ and η are semimartingales, so the stochastic integral in (12) is well defined without further conditions; in particular, no moment conditions on ξ or η are needed.

One motivation for studying (12) is that special cases of it occupy central positions in certain models of financial time series; the Lévy driven OU processes of Barndorff-Nielsen and Shephard (2001a, 2001b, 2003) and the COGARCH process of Klüppelberg, Lindner and Maller (2004) are recent examples.

The GOU as defined in (12) seems to have been first considered by Carmona, Petit and Yor (1997); it is also implicit in the paper of de Haan and

Karandikar (1989), where it occurs as a natural continuous time generalisation of a random recurrence equation. It has been studied in some detail by Lindner and Maller (2005), with emphasis on carrying over some of the properties enjoyed by the Gaussian OU. Other applications are in option pricing (Yor (1992, 2001)), insurance and perpetuities (Harrison (1977), Dufresne (1990), Paulsen and Hove (1999)), and risk theory (Klüppelberg and Kostadinova (2006)). Many of these place further restrictions on ξ and η; for example, ξ may be independent of η, or one or another or both of ξ or η may be a Brownian motion or compound Poisson process, etc. To begin with, we make no assumptions on ξ or η (not even independence), and investigate some general properties of X_t.

Thus, it is the case that X_t is a time homogeneous Markov process (Carmona et al. 1997, Lemma 5.1), and it is elementary that $(X_t)_{t\geq 0}$ is strictly stationary if and only if X_t converges in distribution to X_0, as $t \to \infty$. To study when this occurs, stationarity is related to the convergence of a certain stochastic integral in Lindner and Maller (2005). But which integral? Let us note that in general there is no counterpart of the equality (in distribution) of (2) and (6). That is, in general, $e^{-\xi_t}\int_0^t e^{\xi_{s-}}\mathrm{d}\eta_s$ does not have the same distribution (even for a fixed $t > 0$) as $\int_0^t e^{-\xi_{s-}}\mathrm{d}\eta_s$, as might at first be thought via a time-reversal argument. The correct relationship is given in Proposition 2.3 of Lindner and Maller (2005):

$$e^{-\xi_t}\int_0^t e^{\xi_{s-}}\mathrm{d}\eta_s \stackrel{D}{=} \int_0^t e^{-\xi_{s-}}\mathrm{d}L_s, \text{ for each } t > 0, \tag{13}$$

where L_t is a Lévy process constructed from ξ and η as follows:

$$L_t := \eta_t + \sum_{0<s\leq t}(e^{-\Delta\xi_s} - 1)\Delta\eta_s - t\,\mathrm{Cov}(B_{\xi,1}, B_{\eta,1}), \quad t \geq 0. \tag{14}$$

Here "Cov" denotes the covariance of the Brownian components of ξ and η. In general $L_t \neq \eta_t$, but when ξ and η are independent, for example, they have no jumps in common, a.s., and the covariance term is 0, so (14) gives $L_t \equiv \eta_t$, and the integral on the righthand side of (13) then equals $\int_0^t e^{-\xi_{s-}}\mathrm{d}\eta_s$.

But even in the general case, (13) can be used to investigate the large time behaviour of X_t, because necessary and sufficient conditions for the convergence (a.s., or, in distribution) of Lévy integrals of the form $\int_0^\infty e^{-\xi_{t-}}\mathrm{d}L_t$ have been worked out by Erickson and Maller (2004), phrased in terms of quite simple functionals of the canonical triplet of (L, η), which is easily obtained from the canonical triplet of (ξ, η) via (14). Except for a degenerate case, necessary and sufficient is that $\lim_{t\to\infty}\xi_t = \infty$ a.s., together with a kind of log-moment condition involving only the *marginal* measures of ξ and η. The divergence criterion $\lim_{t\to\infty}\xi_t = \infty$ a.s. is also easily expressed in terms of the canonical measure of ξ_t. The stationarity criterion, given in Theorem 2.1 of Lindner and Maller (2005), is that $(X_t)_{t\geq 0}$ is strictly stationary, for an

appropriate choice of X_0, if and only if the integral $\int_0^\infty e^{-\xi_{t-}}\mathrm{d}L_t$ converges (a.s., or, equivalently, in distribution), or else X_t is indistinguishable from a constant process.

From these results we see that a study of the GOU process can be reduced in part to a study of the exponential Lévy integral $\int_0^\infty e^{-\xi_{t-}}\mathrm{d}L_t$, and this program is continued in Erickson and Maller (2007) (conditions for convergence of stochastic integrals), Bertoin, Lindner and Maller (2008) and Kondo, Maejima and Sato (2006) (continuity properties of the integral), and Maller, Müller and Szimayer (2008) (discrete approximation and statistical properties).

We took as starting point in this section a generalisation of (2), via (12). (12) has direct relevance to stochastic volatility and other models in finance, among other possible applications. On the other hand, modelling by SDEs such as (1) (the Langevin equation) can arise directly from a physical situation; e.g., the interpretation of (1) as describing the motion of a particle under a restraining force proportional to its velocity. The counterpart of (1) for the GOU is the SDE

$$\mathrm{d}X_t = (X_{t-} - m)\mathrm{d}U_t + \mathrm{d}L_t,\ t \geq 0, \tag{15}$$

where (U, L) is a bivariate Lévy process. Suppose this holds for a U whose Lévy measure attributes no mass to $(-\infty, -1]$, and define a Lévy process ξ by $\xi_t = -\log \mathcal{E}(U)_t$, where $\mathcal{E}(U)$ denotes the *stochastic exponential* of U, namely, the solution to the SDE $\mathrm{d}\mathcal{E}(U)_t = \mathcal{E}(U)_{t-}\mathrm{d}U_t$ with $\mathcal{E}(U)_0 = 1$; see Protter (2005, p. 85). Then define a Lévy process η_t by

$$\eta_t := L_t - \sum_{0<s\leq t}(1 - e^{-\Delta\xi_s})\Delta L_s + t\,\mathrm{Cov}(B_{\xi,1}, B_{L,1}), \quad t \geq 0. \tag{16}$$

With these definitions, (12) is the unique (up to indistinguishability) solution to (15). To verify this, use integration by parts in (12) together with Eq. (2.10) of Lindner and Maller (2005). The fact that the Lévy measure of U attributes no mass to $(-\infty, -1]$ ensures that $\mathcal{E}(U)$ is positive. Conversely, if, for a given bivariate Lévy process (ξ, η), L satisfies (14), and U satisfies $\xi_t = -\log \mathcal{E}(U)_t$, then X_t as defined in (12) satisfies (15), and, further, the Lévy measure of U attributes no mass to $(-\infty, -1]$. See Protter (2005, p. 329) and Yoeurp (1979) for further discussion.

A third approach to generalising an OU is to consider more general time changes. Monroe (1978), generalising Lévy's result for continuous local martingales, showed that *any* semimartingale can be obtained as a time changed Brownian motion. Thus we can write $\int_0^t e^{\xi_{s-}}\mathrm{d}\eta_s = W_{T_t}$, for an SBM W and an increasing semimartingale $(T_t)_{t\geq 0}$, leading to another kind of generalisation of (3). The properties of such a class are also unexplored, so far as we know. Other versions of time changed Brownian motions have been used in many situations; see, e.g., Anh, Heyde and Leonenko (2002), for a financial application.

3.3 Self-decomposability, self-similarity, class L, Lamperti transform

Consider the case when $m = 0$ and $\xi_t = \gamma t$, $\gamma > 0$, is a pure drift in (12):

$$X_t = e^{-\gamma t} \int_0^t e^{\gamma s} \mathrm{d}\eta_s + X_0 e^{-\gamma t},\ t \geq 0. \tag{17}$$

Say that (the distribution of) a random variable X is *semi-self-decomposable* if X has the same distribution as $aX + Y^{(a)}$, for a constant $0 < a < 1$, for some random variable $Y^{(a)}$, independent of X, possibly depending on a. If an equality in distribution $X \stackrel{D}{=} aX + Y^{(a)}$ can be achieved for *all* $a \in (0, 1)$, X is said to be *self-decomposable*. See Sato (1999, Section 15). This property can also be described as saying that the distribution of X is *of Class L*; this is a subclass of the infinitely divisible distributions which can be obtained as the limit laws of normed, centered, sums of independent (but not necessarily identically distributed) random variables. See Feller (1971, p. 588). Class L contains but is not confined to the stable laws, which are the limit laws of normed, centered, sums of i.i.d. random variables.

A potential limiting value of X_t in (17) as $t \to \infty$ is the random variable $X_\infty := \int_0^\infty e^{-\gamma t} \mathrm{d}\eta_t$, if finite, and then X_t is stationary if $X_0 \stackrel{D}{=} X_\infty$. Wolfe (1982) showed that a random variable X is self-decomposable if and only if it has the representation

$$X \stackrel{D}{=} \int_0^\infty e^{-\gamma t} \mathrm{d}\eta_t,$$

for some Lévy process η with $E \log^+ |\eta| < \infty$ (and then X is a.s. finite), and, further, that the canonical triplets of X_t (the Lévy process with the distribution of X when $t = 1$) and η_t are then connected in a simple way. He made crucial use of the formula

$$E e^{\mathrm{i}\theta \int_a^b f(s) \mathrm{d}\eta_s} = E e^{\int_a^b \Psi_\eta(-\theta f(s)) \mathrm{d}s},\ 0 \leq a < b < \infty, \theta \in \mathbb{R}, \tag{18}$$

where f is a bounded continuous function in $\mathbb{R}$ and $\Psi_\eta(\theta) := -\log\left(E e^{\mathrm{i}\theta \eta_1}\right)$ (e.g., Bichteler (2002, Lemma 4.6.4, p. 256)).

An *H–self-similar* process $(X_t)_{t \geq 0}$ is such that $(X_{at})_{t \geq 0}$ has the same distribution as $(a^H X_t)_{t \geq 0}$, for some constant $H > 0$, and each $a > 0$. Sato (1991) showed that a random variable X_1 is self-decomposable if and only if for each $H > 0$ its distribution is the distribution at time 1 of an H–self-similar process. An H–self-similar *Lévy* process must have $H \geq 1/2$; and then X_t is an α–stable process with index $\alpha = 1/H \in (0, 2]$.

The *Lamperti Transform* of an H–self-similar process $(X_t)_{t \geq 0}$ is the (stationary) process $Y_t := e^{-tH} X_{e^t}$, $t \geq 0$. Lamperti (1962, 1972) showed, conversely, that any stationary process Y can be represented in this form. Thus, in summary, we have a correspondence between a stationary process Y_t, an

H–self-similar process X_t, a self-decomposable random variable X_1, the class L, and the integral $\int_0^\infty e^{-\gamma t}\mathrm{d}\eta_t$. Jeanblanc, Pitman and Yor (2002) give an elegant linking approach to these.

The integral $\int_0^\infty e^{-\xi_t}\mathrm{d}t$ (assumed convergent) is self-decomposable when ξ is spectrally negative, but *not* in general; (in fact, it is not even infinitely divisible in general). These results are due to Samorodnitsky (reported in Klüppelberg et al. (2004)). Thus, *a fortiori*, the integral $\int_0^\infty e^{-\xi_t}\mathrm{d}\eta_t$ is not in general self-decomposable. See also Kondo et al. (2006) for further results.

4 Discretisations

4.1 Autoregressive representation, and perpetuities

Given a Lévy process L_t with $E\log^+|L_1| < \infty$ and constants $h > 0$ and $\gamma > 0$, let $(Q_n)_{n=1,2,\ldots}$ be i.i.d. with the distribution of $e^{-\gamma h}\int_0^h e^{\gamma s}\mathrm{d}L_s$. Then (Wolfe (1982)) the discrete time process (time series) defined recursively by

$$Z_n = e^{-\gamma h}Z_{n-1} + Q_n,\ n = 1, 2, \ldots, \quad \text{with } Z_0 = 0, \tag{19}$$

converges in distribution as $n \to \infty$ to a random variable with the distribution of the (a.s. finite) integral $\int_0^\infty e^{-\gamma t}\mathrm{d}L_t$. Thus the stationary distribution of an OU process driven by Lévy motion can be obtained from the behaviour at large times of an autoregressive time series. Conversely, Wolfe (1982) showed that if $(Q_n)_{n=1,2,\ldots}$ are given i.i.d. random variables with $E\log^+|Q_n| < \infty$, and Z_n are defined by the recursion in (19) with $\gamma > 0$ and $h = 1$, then there is a Lévy process L_t with $E\log^+|L_1| < \infty$ such that the process X_t as defined in (17) satisfies $X_n = Z_n$, $n = 1, 2, \ldots$, if and only if $Q_1 \stackrel{D}{=} e^{-\gamma}\int_0^1 e^{\gamma s}\mathrm{d}L_s$. He further gave necessary and sufficient conditions for the latter property to hold; namely, a random variable Q has the same distribution as $e^{-\gamma}\int_0^1 e^{\gamma s}\mathrm{d}L_s$, for a given $\gamma > 0$ and Lévy process L_t with $E\log^+|L_1| < \infty$, if and only if $\prod_{j=0}^\infty E(e^{\mathrm{i}\rho^j\theta Q})$ is the characteristic function of a distribution in class L, where $\rho = e^{-\gamma}$. See also Sato (1999, Section 17).

(19) is a special case of a discrete time "perpetuity". More generally, we may replace the coefficient $e^{-\gamma h}$ in (19) by a random sequence, M_n, say, such that $(Q_n, M_n)_{n=1,2,\ldots}$ is an i.i.d. sequence of 2-vectors. Then Z_n is a kind of analogue of the Lévy integral $\int_0^t e^{-\xi_{s-}}d\eta_s$; see, e.g., Lindner and Maller (2005) for a discussion. Random sequences related to perpetuities have received much attention in the literature as models for a great variety of phenomena, including but not restricted to the actuarial area. We refer to Vervaat (1979), Goldie and Maller (2000), Nyrhinen (1999, 2001).

4.2 Statistical issues: Estimation and hypothesis testing

There are methods of estimation of parameters in continuous time models based on hypothetical continuous time observation of a process over a finite interval, and the testing of hypotheses about them, for example, in a likelihood framework (Liptser and Shiryaev (1978), Basawa and Prakasa Rao (1980), Heyde (1997), Kutoyants (2004)), which provide much insight. But in practise we can only observe in discrete time, and have to think how to approximate the parameters in the original continuous time model from a finite (discrete) sample. Furthermore, observation in practise can only be carried out over a finite time interval, whereas frequently in statistics we may wish to employ a large sample theory, or, in the case of a time series, let the observation time grow large, to derive benchmark distributions for parameter estimates and test statistics which are free from finite sample effects.

Consequently, in approximating a continuous by a discrete time process, we can proceed in one or both of two ways. One is to form a series of approximations on a finite time interval $[0, T]$, which is subdivided into finer and finer grids, so that in the limit the discrete approximations converge, hopefully, to the original continuous, time process (in some mode); alternatively, we can sample at discrete points in a finite time interval $[0, T]$ and let $T \to \infty$ to get asymptotic distributions; or, thirdly, we can attempt to combine both methods in some way.

4.3 Discretely sampled process

Discrete sampling of an OU process on an equispaced grid over a finite time horizon $T > 0$ produces an autoregressive (AR) time series, as follows. Suppose X_t satisfies (1), and fix a compact interval $[0, T]$, $T > 0$. Then

$$X_{i,n} = X_{iT/n},\ i = 0, 1, \ldots, n,\ \text{for } n = 1, 2, \ldots, \tag{20}$$

is the discretely sampled process. From (1) we can write

$$X_{i,n} = (1 - \alpha_n)m + \alpha_n X_{i-1,n} + \sigma_n \varepsilon_{i,n},\ i = 1, 2, \ldots, n, \tag{21}$$

where

$$\alpha_n = e^{-\gamma T/n}, \quad \sigma_n^2 = \sigma^2(1 - e^{-2\gamma T/n})/(2\gamma), \tag{22}$$

and

$$\varepsilon_{i,n} := \frac{\sigma}{\sigma_n} \int_0^{T/n} e^{\gamma(s - T/n)} \mathrm{d}B_{s+(i-1)T/n}. \tag{23}$$

(21) is a system of autoregressions, where the $(\varepsilon_{i,n})_{i=1,2,\ldots,n}$ are i.i.d. standard normal random variables for each $n = 1, 2, \ldots$.

Next, embed each $X_{i,n}$ into a continuous time process $X_n(t)$ by setting

$$X_n(t) = X_{i-1,n}, \text{ for } (i-1)T/n \le t < iT/n, \ i = 1, 2, \ldots, n. \tag{24}$$

Then $X_n(t) \to X_t$, uniformly on $[0, T]$, in probability, as $n \to \infty$.

Szimayer and Maller (2004) carry out the above procedure, but with a Lévy process L_t, satisfying $EL_1 = 0$ and $EL_1^2 = 1$, replacing B_t in (1) and consequently in (23). The $\varepsilon_{i,n}$ in (23) remain i.i.d. $(0, 1)$ random variables, though in general of course they are no longer normally distributed. Szimayer and Maller (2004) used a Quasi-Maximum Likelihood (QML) approach, whereby a likelihood for the observations is written down as if the $\varepsilon_{i,n}$ were normally distributed, and estimates and test statistics calculated from it, but then the normality assumption is discarded for the rest of the analysis. They test the hypothesis $H_0 : \gamma = 0$, of no mean reversion in the model (so X_t reduces to L_t, a pure Lévy process). This hypothesis test has the nonstandard feature that the long term equilibrium parameter m "disappears under the null"; it cannot be identified from (1) when $\gamma = 0$. Methods of Davies (1977, 1987) are available for handling this. Szimayer and Maller (2004) work out the asymptotic distribution (as the mesh size tends to 0, over the compact interval $[0, T]$) of the QML statistic for testing H_0, as a function of the underlying Lévy process L_t. That asymptotic distribution of course depends on T, and as $T \to \infty$, Szimayer and Maller (2004) show further that it tends to the distribution of a random variable related to the Dickey-Fuller unit root test in econometrics. This procedure is an example of estimating on a finite grid whose mesh size shrinks to 0, after which the observation window expands to infinity.

4.4 Approximating the COGARCH

The COGARCH is a continuous-time dynamic model suggested by Klüppelberg, Lindner and Maller (2004) to generalise the popular GARCH (Generalised Autoregressive Conditional Heteroscedasticity) model now commonly used in (discrete) time series analysis. The COGARCH is defined by

$$G_t = \int_0^t \sigma_{s-} \mathrm{d}L_s, \ t \ge 0, \tag{25}$$

where L_t is a "background driving Lévy process", and σ_t, the volatility process, satisfies

$$\sigma_t^2 = \beta e^{-X_t} \int_0^t e^{X_s} \, \mathrm{d}s + \sigma_0^2 e^{-X_t}, \ t \ge 0, \tag{26}$$

for constants $\beta > 0$ and $\sigma_0^2 > 0$. (26) is a version of the GOU (12), with η_t replaced by a pure drift, and ξ_t replaced by X_t. The latter is just a notational

change; the X_t in (26) is also a Lévy process, defined in terms of the original L_t by

$$X_t = \eta t - \sum_{0<s\le t} \log(1+\varphi(\Delta L_s)^2),\ t \ge 0, \tag{27}$$

for parameters $\eta > 0$ and $\varphi > 0$. Note that only one source of randomness, L_t, underlies both the process itself and the volatility process; this is an important feature of the discrete time GARCH models, preserved in the COGARCH.

Further analysis of the COGARCH is in Klüppelberg et al. (2004, 2006), where stationarity properties are related to the convergence of a Lévy integral. See also Lindner (2008). Statistical issues, especially, fitting the COGARCH to data, are in Haug et al. (2007), Müller (2007), and Maller, Müller and Szimayer (2008). The latter paper proposes a discretisation of the COGARCH in the same spirit as we discussed above for the Lévy driven OU model. Using a first-jump approximation of a Lévy process originally developed in Szimayer and Maller (2007) for an option pricing application, Maller et al. (2008) show that the COGARCH can be obtained as a limit of discrete time GARCH processes defined on the same probability space. This allows advantage to be taken of currently existing methods in time-series modeling and econometrics for this well-established process class.

The procedure is as follows. Take a sequence of integers $(N_n)_{n\ge1}$ with $\lim_{n\to\infty} N_n = \infty$, and a finite interval $[0,T]$, $T > 0$, with a deterministic partitioning $0 = t_0(n) < t_1(n) < \ldots < t_{N_n}(n) = T$. Let $\Delta t_i(n) := t_i(n) - t_{i-1}(n)$ for $i = 1,2,\ldots,N_n$, and assume $\Delta t_n := \max_{i=1,\ldots,N_n} \Delta t_i(n) \to 0$ as $n \to \infty$. Given the COGARCH parameters (β,η,φ), define the process

$$G_{i,n} = G_{i-1,n} + \sigma_{i-1,n}\sqrt{\Delta t_i(n)}\varepsilon_{i,n},\ \text{for } i = 1,2,\ldots,N_n,\ \text{with } G_{0,n} = 0, \tag{28}$$

with an accompanying variance process:

$$\sigma^2_{i,n} = \beta\Delta t_i(n) + \left(1+\varphi\Delta t_i(n)\varepsilon^2_{i,n}\right) e^{-\eta\Delta t_i(n)}\sigma^2_{i-1,n},\ i = 1,2,\ldots,N_n. \tag{29}$$

Here, for each $n \ge 1$, $(\varepsilon_{i,n})_{i=1,\ldots,N_n}$ is a sequence of independent random variables with $\mathbb{E}\varepsilon_{1,n} = 0$ and $\mathbb{E}\varepsilon^2_{1,n} = 1$ constructed pathwise from the driving Lévy process L_t in (25) and its characteristics; and $\sigma^2_{0,n}$ is a given random variable, independent of the $\varepsilon_{i,n}$. (28) and (29) define a kind of discrete time GARCH-type recursion with scaling by the time increments $\Delta t_i(n)$.

The discrete time processes are then embedded into continuous time by

$$G_n(t) := G_{i,n} \quad \text{and} \quad \sigma^2_n(t) := \sigma^2_{i,n} \text{ when } t \in (t_{i-1}(n), t_i(n)],\ 0 \le t \le T, \tag{30}$$

with $G_n(0) = 0$ and $\sigma^2_n(0) = \sigma^2_{0,n}$. A key result of Maller et al. (2008) is that, as $n \to \infty$ (so $\Delta t(n) \to 0$), the Skorokhod distance between $(G_n(\cdot),\sigma_n(\cdot))$ and $(G(\cdot),\sigma(\cdot))$, over $[0,T]$, converges in probability to 0; thus, in particular, $(G_n(\cdot),\sigma_n(\cdot))$ converges in distribution to $(G(\cdot),\sigma(\cdot))$ in $\mathbb{D}[0,T]\times\mathbb{D}[0,T]$, where $\mathbb{D}[0,T]$ is the space of càdlàg stochastic process on $[0,T]$.

Maller et al. (2008) use this result to motivate an estimation procedure for the COGARCH parameters in terms of estimates of the parameters of the discrete GARCH approximating process. Via some simulations, this is shown to work somewhat better, in some selected situations, than the Haug et al. (2007) method, at least as judged by the mean square error of the estimates.

As an example application, we fitted the COGARCH model to a series of 33,480 log-prices of the Intel stock traded on the NYSE between February 1 and June 6, 2002, observed every minute from 09:36am to 04:00pm. The data is from the TAQ data base provided by the NYSE. We removed the overnight jumps and a linear trend from the data, then fitted a GARCH model by the QML method as described above, thus obtaining estimates $(\widehat{\beta}, \widehat{\varphi}, \widehat{\eta})$ of (β, φ, η). Then with G_t as the log stock price at time t, an estimate of the volatility process $(\sigma_t^2)_{t\geq 0}$ can be calculated recursively from

$$\widehat{\sigma}_n^2 = \widehat{\beta} + (1 - \widehat{\eta})\widehat{\sigma}_{n-1}^2 + \widehat{\varphi}(G_n - G_{n-1})^2, \ n = 1, 2, \ldots$$

(Haug et al. 2007). Figure 1 shows that the resulting volatility sequence (for the first 1,000 observations) compares reasonably well with the absolute log returns. Further discussion is in Maller et al. (2008).

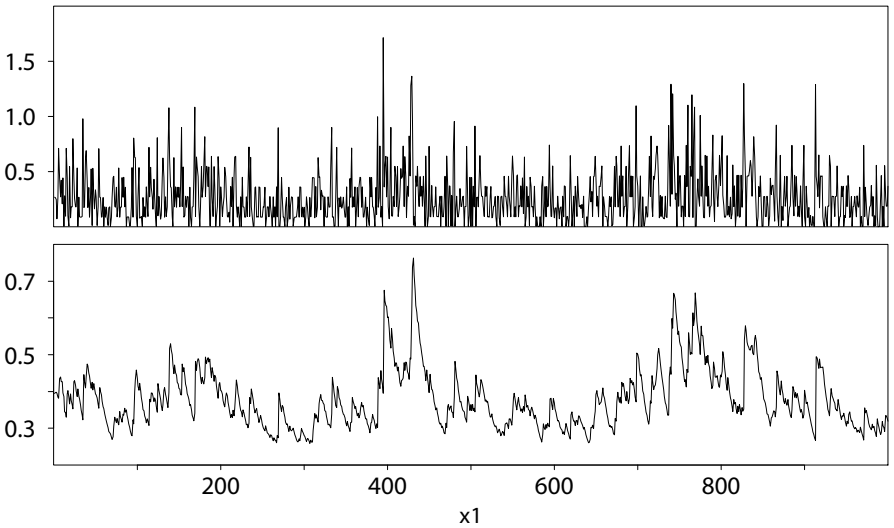

Fig. 1 Top: 1,000 minute-by-minute absolute log returns on Intel stock. Bottom: Corresponding estimated annualized volatilities for Intel data.

5 Conclusion

This survey cannot do justice in the space available to the many theoretical and practical studies, past and ongoing, related to the OU and GOU processes. We note in particular Brockwell, Erdenebaatar and Lindner (2006) (a COGARCH(p, q)); Masuda (2004) (a multidimensional OU process); Aalen and Gjessing (2004) (an interesting connection between the finance and survival analysis applications); Novikov (2004) (passage time problems); Kondo et al. (2006) (multidimensional exponential Lévy integrals); and the list goes on. Despite all this activity, much remains to be done, as we have suggested throughout the discussion, to add to our understanding of the stochastic processes themselves, and their statistical properties.

Acknowledgement We thank Peter Brockwell, Claudia Klüppelberg, Vicky Fasen and Alex Lindner for some helpful comments; and the latter two, in particular, for discussions relating to the SDE in (15). This research was partially supported by ARC grant DP0664603.

References

Aalen, O. O. and Gjessing, H. K. (2004): Survival models based on the Ornstein-Uhlenbeck process. *Lifetime Data Analysis* **10**, 407–423.

Anh, V. V., Heyde, C. C. and Leonenko, N. N. (2002): Dynamic models of long-memory processes driven by Lévy noise. *J. Appl. Probab.* **39**, 730–747.

Barndorff-Nielsen, O. E. and Shephard, N. (2001a): Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion). *J. Roy. Statist. Soc. Ser. B* **63**, 167–241.

Barndorff-Nielsen, O. E. and Shephard, N. (2001b): Modelling by Lévy processes for financial econometrics. *In: Barndorff-Nielsen, O. E., Mikosch, T. and Resnick S. (eds:) Lévy processes, theory and applications.* Birkhäuser, Boston.

Barndorff-Nielsen, O. E. and Shephard, N. (2003): Integrated OU processes and non-Gaussian OU-based stochastic volatility models. *Scand. J. Statist.* **30**, 277–295.

Basawa, I. V. and Prakasa Rao B. L. S. (1980): *Statistical inference for stochastic processes.* Academic Press, London New York.

Bertoin, J. (1996): *Lévy processes.* Cambridge University Press, Cambridge.

Bichteler, K. (2002): *Stochastic integration with jumps.* Cambridge University Press, Cambridge.

Bertoin, J., Lindner, A. and Maller, R.A. (2008): *On continuity properties of the law of integrals of Lévy processes. Séminaire de Probabilités XLI, Lecture Notes in Mathematics* **1934**. Springer, Heidelberg.

Brockwell, P. J. (2008): Lévy-driven continuous time ARMA processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 456–480. Springer, New York.

Brockwell, P., Erdenebaatar, C. and Lindner, A. (2006): Continuous time GARCH processes. *Annals of Appl. Prob.* **16**, 790–826.

Carmona, P., Petit, F. and Yor, M. (1997): On the distribution and asymptotic results for exponential functionals of Lévy processes. *In: Yor, M. (Ed.): Exponential functionals and principal values related to Brownian motion.* Bibl. Rev. Mat. Iberoamericana, Madrid.

Davies, R. B. (1977): Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.

Davies, R. B. (1987): Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

De Haan, L. and Karandikar, R. L. (1989): Embedding a stochastic difference equation in a continuous-time process. *Stoch. Proc. Appl.* **32**, 225–235.

Dufresne, D. (1990): The distribution of a perpetuity, with application to risk theory and pension funding. *Scand. Actuar. J.* **9**, 39–79.

Erickson, K. B. and Maller, R. A. (2004): Generalised Ornstein-Uhlenbeck processes and the convergence of Lévy integrals, *Séminaire de Probabilités XXXVIII, Lecture Notes in Mathematics* **1857**, 70–94. Springer-Verlag Heidelberg.

Erickson, K. B. and Maller, R. A. (2007): Finiteness of functions of integrals of Lévy processes, *Proc. Lond. Math. Soc.* **94**, 386–420.

Feller, W. (1971): *An introduction to probability theory and its applications II.* Wiley, New York.

Geman, H., Madan. D. B. and Yor, M. (2000): Asset prices are Brownian motion: only in business time. *In: Avellaneda, M. (Ed.): Quantitative analysis in financial markets* **2**. World Scientific Publishing Company, Singapore.

Goldie, C. M. and Maller, R. A. (2000): Stability of perpetuities. *Ann. Probab.* **28**, 1195–1218.

Harrison, J. M. (1977): Ruin problems with compounding assets. *Stoch. Proc. Appl.* **5**, 67–79.

Haug, S., Klüppelberg, C., Lindner, A. and Zapp, M. (2007): Method of moments estimation in the COGARCH(1,1) model. *The Econ. J.* **10**, 320–341.

Heyde, C. C. (1997): *Quasi-likelihood and its application: a general approach to optimal parameter estimation.* Springer, Berlin Heidelberg New York.

Jeanblanc, M., Pitman, J. and Yor, M. (2002): Self-similar processes with independent increments associated with Lévy and Bessel processes. *Stoch. Proc. Appl.* **100**, 223–231.

Karatzas, I. and Shreve, S. E. (1998): *Methods of mathematical finance.* Springer, Berlin Heidelberg New York.

Klüppelberg, C. and Kostadinova, R. (2006): Integrated insurance risk models with exponential Lévy investment. *Insurance: Math. and Econ.* to appear.

Klüppelberg, C., Lindner, A. and Maller, R. A. (2004): A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour. *J. Appl. Prob.* **41**, 601–622.

Klüppelberg, C., Lindner, A. and Maller, R. A. (2006): Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models. *In: Kabanov, Y., Liptser, R. and Stoyanov, J. (Eds.): From stochastic calculus to mathematical finance, Shiryaev Festschrift.* Springer, Berlin Heidelberg New York.

Kondo, H., Maejima, M. and Sato, K. (2006): Some properties of exponential integrals of Lévy processes and examples. *Elect. Comm. in Probab.* **11**, 291–303. (www.math.keio.ac.jp/local/maejima/).

Kutoyants, Y. A. (2004): *Statistical inference for ergodic diffusion processes.* Springer, Berlin Heidelberg New York.

Lamperti, J. (1962): Semi-stable stochastic processes. *Trans. Amer. Math. Soc.* **104**, 62–78.

Lamperti, J. (1972): Semi-stable Markov processes. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **22**, 205–225.

Lindner, A. (2008): Continuous time approximations to GARCH and stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 481–496. Springer, New York.

Lindner, A. and Maller, R. A. (2005): Lévy integrals and the stationarity of generalised Ornstein-Uhlenbeck processes. *Stoch. Proc. Appl.* **115**, 1701–1722.

Liptser, R. S. and Shiryaev, A. N. (1978): *Statistics of random processes II: applications.* Springer, Berlin Heidelberg New York.

Madan, D. B., Carr, P. P. and Chang, E. C. (1998): The Variance Gamma process and option pricing. *European Finance Review* **2**, 79–105.

Maller, R.A., Müller, G. and Szimayer, A. (2008): GARCH modelling in continuous time for irregularly spaced time series data. *Bernoulli* **14**, 519–542.

Masuda, H. (2004): On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process. *Bernoulli* **10**, 97–120.

Monroe, I. (1978): Processes that can be embedded in Brownian motion. *Ann. Probab.* **6**, 42–56.

Müller, G. (2007): MCMC estimation of the COGARCH(1,1) model. *Preprint, Munich University of Technology* (http://www-m4.ma.tum.de/Papers/).

Novikov, A. A. (2004): Martingales and first-exit times for the Ornstein-Uhlenbeck process with jumps. *Theory Probab. Appl.* **48**, 288–303.

Nyrhinen, H. (1999): On the ruin probabilities in a general economic environment. *Stoch. Proc. Appl.* **83**, 319–330.

Nyrhinen, H. (2001): Finite and infinite time ruin probabilities in a stochastic economic environment. *Stoch. Proc. Appl.* **92**, 265–285.

Paulsen, J. and Hove, A. (1999): Markov chain Monte Carlo simulation of the distribution of some perpetuities. *Adv. Appl. Prob.* **31**, 112–134.

Protter, P. (2005): *Stochastic integration and differential equations.* 2nd edition version 2.1. Springer, Berlin Heidelberg New York.

Sato, K. (1991): Self-similar processes with independent increments. *Prob. Th. Rel. Fields* **89**, 285–300.

Sato, K. (1999): Lévy processes and infinitely divisible distributions. Cambridge University Press, Cambridge.

Szimayer, A. and Maller, R. A. (2004): Testing for mean-reversion in processes of Ornstein-Uhlenbeck type. *Stat. Inf. for Stoch. Proc.* **7**, 95-113.

Szimayer, A. and Maller, R. A. (2007): Finite approximation schemes for Lévy processes, and their application to optimal stopping problems. *Stoch. Proc. Appl.* **117**, 1422–1447. to appear.

Uhlenbeck, G. E. and Ornstein, L. S. (1930): On the theory of the Brownian motion. *Phys. Rev.* **36**, 823–841.

Vasicek, O. A. (1977): An equilibrium characterisation of the term structure. *J. Fin. Econ.* **5**, 177–188.

Vervaat, W. (1979): On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Adv. Appl. Prob.* **11**, 750–783.

Wolfe, S. J. (1982): On a continuous analogue of the stochastic difference equation $X_n = \rho X_{n-1} + B_n$. *Stoch. Proc. Appl.* **12**, 301–312.

Yoeurp, Ch. (1979): Solution explicite de l'équation $Z_t = 1 + \int_0^t |Z_{s-}| dX_s$. *Séminaire de Probabilités X, Lecture Notes in Mathematics, XXXVIII* **511**, 614–619. Springer-Verlag, Heidelberg.

Yor, M. (1992): Sur certaines fonctionnelles du mouvement brownien réel. *J. Appl. Probab.* **29**, 202–208.

Yor, M. (2001): *Exponential functionals of Brownian motion and related processes.* Springer, Berlin Heidelberg New York.

Jump–Type Lévy Processes

Ernst Eberlein

Abstract Lévy processes are developed in the more general framework of semimartingale theory with a focus on purely discontinuous processes. The fundamental exponential Lévy model is given, which allows us to describe stock prices or indices in a more realistic way than classical diffusion models. A number of standard examples including generalized hyperbolic and CGMY Lévy processes are considered in detail.

1 Probabilistic Structure of Lévy Processes

The assumption that observations are normally distributed is predominant in many areas of statistics. So is the situation with time series of financial data, where from the very beginning of continuous-time modeling, Brownian motion itself or geometric Brownian motion became the favorite. This is largely due to the fact that the normal distribution as well as the continuous-time process it generates have nice analytic properties. The standard techniques to handle these objects are known to a large community, which at the same time is less familiar with more sophisticated distributions and processes. On the other hand, a thorough look at data from various areas of finance, such as equity, fixed income, foreign exchange or credit, clearly reveals that assuming normality, one gets a model which is only a poor approximation of reality. If $(S_t)_{t\geq 0}$ denotes a price process in continuous or discrete time, the quantity to be considered is the *log returns*:

$$\log S_{t+\delta} - \log S_t = \log(S_{t+\delta}/S_t). \tag{1}$$

Ernst Eberlein
Department of Mathematical Stochastics, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany, e-mail: eberlein@stochastik.uni-freiburg.de

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_19,

Usually log returns are preferred to *relative price changes* $(S_{t+\delta} - S_t)/S_t$ because by adding up log returns over n periods, one gets the log return for the period $n\delta$. This is not the case for relative price changes. Numerically, the difference between log returns and relative price changes is negligible because $x - 1$ is well approximated by $\log x$ at $x = 1$.

Whereas log returns taken from monthly stock prices (S_t) are reasonably represented by a normal distribution, the deviation becomes significant if one considers prices on a daily or even an intraday time grid (Eberlein and Keller (1995), Eberlein and Özkan (2003b)). As a consequence of the high volumes traded nowadays on electronic platforms, daily price changes of several percent are rather frequent also for big companies, i.e., companies with a high market capitalization. A model based on the Gaussian distribution however would allow this order of price change only as a very rare event. Let us underline that the deviation of probabilities is not restricted to tails only, but can be observed on a lower scale for small price movements as well. Empirical return distributions have substantially more mass around the origin than the normal distribution. In order to improve the statistical accuracy of the models and thus to improve derivative pricing, risk management and portfolio optimization to name just some key areas of application, many extensions of the basic models have been introduced. Let us refer to adding stochastic volatility, stochastic interest rates, correlation terms and so on. Without any doubt these extensions typically reduce the deviation between model and reality. On the other hand, in most cases the simple analytic properties are sacrificed and in particular the distributions which the extended models produce are no longer known explicitly.

A more fundamental change in the modeling approach is to consider from the very beginning more realistic distributions and to keep the analytic form of the model itself simple. This leads naturally to a broader class of driving processes, namely, Lévy processes. A *Lévy process* $X = (X_t)_{t\geq 0}$ is a process with *stationary* and *independent increments*. Underlying it is a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ to which the process $(X_t)_{t\geq 0}$ is adapted. It is well known (Theorem 30 in Protter (2004)) that a Lévy process has a version with *càdlàg paths*, i.e., paths which are right-continuous and have limits from the left. In the following we shall always consider processes with càdlàg paths. A (one-dimensional) Lévy process can be represented in the following way, where we assume $X_0 = 0$ for convenience:

$$X_t = bt + \sqrt{c}W_t + Z_t + \sum_{s\leq t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}}. \tag{2}$$

Here b and $c \geq 0$ are real numbers, $(W_t)_{t\geq 0}$ is a standard Brownian motion and $(Z_t)_{t\geq 0}$ is a purely discontinuous martingale which is independent of $(W_t)_{t\geq 0}$. $\Delta X_s := X_s - X_{s-}$ denotes the jump at time s if there is any and thus the last sum represents the jumps of the process with absolute jump size bigger than 1.

In the case where $c = 0$, i.e., if the continuous Gaussian part disappears, the process is a *purely discontinuous Lévy process*. As we will see later, many examples which are important for modeling in finance are of this type. Let us also mention that in the general case where both martingales, (W_t) and (Z_t), are present, because of their independence they are orthogonal in a Hilbert-space sense. This fact simplifies the analysis considerably because the two components of the process do not interact. As a consequence the classical formulae known for diffusion processes—for example, Itô's formula—are complemented by a term or terms which come from the jump part of X, but no mixed terms have to be considered.

The decomposition of a Lévy process as given in (2) is known as the *Lévy–Itô decomposition*. At the same time every Lévy process is a *semimartingale* and (2) is the so-called *canonical representation* for semimartingales. For a semimartingale $Y = (Y_t)_{t\geq 0}$, the latter is obtained by the following procedure. One first subtracts from Y the sum of the big jumps, e.g., the jumps with absolute jump size bigger than 1. The remaining process

$$Y_t - \sum_{s\leq t} \Delta Y_s \mathbb{1}_{\{|\Delta Y_s|>1\}} \tag{3}$$

has bounded jumps and therefore is a *special semimartingale* (see I.4.21 and I.4.24 in Jacod and Shiryaev (1987)). Any special semimartingale admits a unique decomposition into a local martingale $M = (M_t)_{t\geq 0}$ and a predictable process with finite variation $V = (V_t)_{t\geq 0}$, i.e., the paths of V have finite variation over each finite interval $[0, t]$. For Lévy processes the finite variation component turns out to be the (deterministic) linear function bt. Any local martingale M (with $M_0 = 0$) admits a unique decomposition (see I.4.18 in Jacod and Shiryaev (1987)) $M = M^c + M^d$, where M^c is a local martingale with continuous paths and M^d is a purely discontinuous local martingale which we denoted Z in (2). For Lévy processes the continuous component M^c is a standard Brownian motion $W = (W_t)_{t\geq 0}$ scaled with a constant factor $\sqrt{c}$.

What we have seen so far is that a Lévy process has two simple components: a linear function and a Brownian motion. Now let us look more carefully at the *jump part*. Because we assumed càdlàg paths, over finite intervals $[0, t]$ any path has only a finite number of jumps with absolute jump size larger than ε for any $\varepsilon > 0$. As a consequence the sum of jumps along $[0, t]$ with absolute jump size bigger than 1 is a finite sum for each path.

Of course instead of the threshold 1, one could use any number $\varepsilon > 0$ here. In contrast to the sum of the big jumps, the sum of the small jumps

$$\sum_{s\leq t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|\leq 1\}} \tag{4}$$

does not converge in general. There are too many small jumps to get convergence. One can force this sum to converge by *compensating* it, i.e., by

subtracting the corresponding average increase of the process along $[0,t]$. The average can be expressed by the intensity $F(dx)$ with which the jumps arrive. More precisely, the following limit exists in the sense of convergence in probability:

$$\lim_{\varepsilon\to 0}\Big(\sum_{s\le t}\Delta X_s \mathbb{1}_{\{\varepsilon\le|\Delta X_s|\le 1\}} - t\int x\mathbb{1}_{\{\varepsilon\le|x|\le 1\}}F(dx)\Big). \tag{5}$$

Note that the first sum represents the (finitely many) jumps of absolute jump size between ε and 1. The integral is the average increase of the process in a unit interval when jumps with absolute size smaller than ε or larger than 1 are eliminated. One cannot separate this difference, because in general neither of the expressions has a finite limit as $\varepsilon\to 0$.

There is a more elegant way to express (5). For this one introduces the *random measure of jumps* of the process X denoted by μ^X:

$$\mu^X(\omega;dt,dx)=\sum_{s>0}\mathbb{1}_{\{\Delta X_s\neq 0\}}\varepsilon_{(s,\Delta X_s(\omega))}(dt,dx). \tag{6}$$

If a path of the process given by ω has a jump of size $\Delta X_s(\omega)=x$ at time point s, then the random measure $\mu^X(\omega;\cdot,\cdot)$ places a unit mass $\varepsilon_{(s,x)}$ at the point (s,x) in $\mathbb{R}_+\times\mathbb{R}$. Consequently for a time interval $[0,t]$ and a set $A\subset\mathbb{R}$, $\mu^X(\omega;[0,t]\times A)$ counts how many jumps of jump size within A occur for this particular path ω from time 0 to t:

$$\mu^X(\omega;[0,t]\times A)=|\{(s,x)\in[0,t]\times A\mid \Delta X_s(\omega)=x\}|. \tag{7}$$

This number is compared with the average number of jumps with size within A. The latter can be expressed by an intensity measure $F(A)$:

$$E\big[\mu^X(\,\cdot\,;[0,t]\times A)\big]=tF(A). \tag{8}$$

With this notation one can write the sum of the big jumps at the end of (2) in the form

$$\int_0^t\int_{\mathbb{R}} x\mathbb{1}_{\{|x|>1\}}\mu^X(ds,dx) \tag{9}$$

and one can express (Z_t), the martingale of compensated jumps of absolute size less than 1, in the form

$$\int_0^t\int_{\mathbb{R}} x\mathbb{1}_{\{|x|\le 1\}}(\mu^X(ds,dx)-dsF(dx)). \tag{10}$$

Note that $\mu^X(\omega,ds,dx)$ is a random measure, i.e., it depends on ω, whereas $dsF(dx)$ is a product measure on $\mathbb{R}_+\times\mathbb{R}$ not depending on ω. Again μ^X and $dsF(dx)$ cannot be separated in general.

2 Distributional Description of Lévy Processes

The distribution of a Lévy process $X = (X_t)_{t>0}$ is completely determined by any of its marginal distributions $\mathcal{L}(X_t)$. Let us consider $\mathcal{L}(X_1)$ and write for any natural number n

$$X_1 = X_{1/n} + (X_{2/n} - X_{1/n}) + (X_{3/n} - X_{2/n}) + \cdots + X_{n/n} + X_{n-1/n}. \quad (11)$$

By stationarity and independence of the increments we see that $\mathcal{L}(X_1)$ is the n-fold convolution of the laws $\mathcal{L}(X_{1/n})$:

$$\mathcal{L}(X_1) = \mathcal{L}(X_{1/n}) * \cdots * \mathcal{L}(X_{1/n}). \quad (12)$$

Consequently $\mathcal{L}(X_1)$ and analogously any $\mathcal{L}(X_t)$ are infinitely divisible distributions. Conversely any infinitely divisible distribution ν generates in a natural way a Lévy process $X = (X_t)_{t\geq 0}$ which is uniquely determined by setting $\mathcal{L}(X_1) = \nu$. If for $n > 0$, ν_n is the probability measure such that $\nu = \nu_n * \cdots * \nu_n$, then one gets immediately for rational time points $t = k/n$ $\mathcal{L}(X_t)$ as the k-fold convolution of ν_n. For irrational time points t, $\mathcal{L}(X_t)$ is determined by a continuity argument (see Chapter 14 in Breiman (1968)). Because the process to be constructed has independent increments, it is sufficient to know the one-dimensional distributions.

As we have seen, the class of infinitely divisible distributions and the class of Lévy processes are in a one-to-one relationship; therefore, if a specific infinitely divisible distribution is characterized by a few parameters the same holds for the corresponding Lévy process. This fact is crucial for the estimation of parameters in financial models which are driven by Lévy processes. The classical example is Brownian motion which is characterized by the parameters μ and σ^2 of the normal distribution $N(\mu, \sigma^2)$. A number of examples which allow more realistic modeling in finance will be considered in the last section.

For an infinitely divisible distribution ν which we can write as $\nu = \mathcal{L}(X_1)$ for a Lévy process $X = (X_t)_{t\geq 0}$, the *Fourier transform* in its *Lévy–Khintchine form* is given by

$$E[\exp(\mathrm{i}uX_1)] = \exp\left[\mathrm{i}ub - \frac{1}{2}u^2c + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}ux} - 1 - \mathrm{i}ux\mathbb{1}_{\{|x|\leq 1\}}\right)F(dx)\right]. \quad (13)$$

The three quantities b, c and F are those which appeared in (2), (8) and (10). They determine the law of X_1, $\mathcal{L}(X_1)$, and thus the process $X = (X_t)_{t\geq 0}$ itself completely. (b, c, F) is called the *Lévy–Khintchine triplet* or in semimartingale terminology the *triplet of local characteristics*. The truncation function $h(x) = x\mathbb{1}_{\{|x|\leq 1\}}$ used in (13) could be replaced by other versions of truncation functions, e.g., smooth functions which are identical to the identity in a neighborhood of the origin and go to zero outside this neighborhood. Changing h results in a different *drift parameter* b, whereas the *diffusion co-*

efficient $c \geq 0$ and the *Lévy measure* F remain unaffected. We note that F does not have mass on 0, $F(\{0\}) = 0$, and satisfies the following integrability condition:

$$\int_{\mathbb{R}} \min(1, x^2) F(dx) < \infty. \tag{14}$$

Conversely any measure on the real line with these two properties together with parameters $b \in \mathbb{R}$ and $c \geq 0$ defines via (13) an infinitely divisible distribution and thus a Lévy process. Let us write (13) in the short form

$$E[\exp(\mathrm{i}uX_1)] = \exp(\psi(u)). \tag{15}$$

ψ is called the *characteristic exponent*. Again by independence and stationarity of the increments of the process (see (11), (12)), one derives that the characteristic function of $\mathcal{L}(X_t)$ is the tth power of the characteristic function of $\mathcal{L}(X_t)$:

$$E[\exp(\mathrm{i}uX_t)] = \exp(t\psi(u)). \tag{16}$$

This property is useful when one has to compute numerically values of derivatives which are represented as expectations of the form $E[f(X_T)]$, where X_T is the value of a Lévy process at maturity T, and the parameters of the Lévy process were estimated as the parameters of $\mathcal{L}(X_1)$.

A lot of information on the process can be derived from *integrability properties* of the Lévy measure F. The following proposition shows that finiteness of moments of the process depends only on the frequency of large jumps since it is related to integration by F over $\{|x| > 1\}$.

Proposition 1 *Let $X = (X_t)_{t \geq 0}$ be a Lévy process with Lévy measure F.*

1. *X_t has finite pth moment for $p \in \mathbb{R}_+$, i.e., $E[|X_t|^p] < \infty$, if and only if $\int_{\{|x|>1\}} |x|^p F(dx) < \infty$.*
2. *X_t has finite pth exponential moment for $p \in \mathbb{R}$, i.e., $E[\exp(pX_t)] < \infty$, if and only if $\int_{\{|x|>1\}} \exp(px) F(dx) < \infty$.*

For the proof see Theorem 25.3 in Sato (1999). From part 1 we see that if the generating distribution $\mathcal{L}(X_1)$ has *finite expectation* then $\int_{\{|x|>1\}} xF(dx) < \infty$. This means that we can add $-\int \mathrm{i}ux \mathbb{1}_{\{|x|>1\}} F(dx)$ to the integral in (13) and get the simpler representation for the Fourier transform:

$$E[\exp(\mathrm{i}uX_1)] = \exp\left[\mathrm{i}ub - \frac{1}{2}u^2 c + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}ux} - 1 - \mathrm{i}ux\right) F(dx)\right]. \tag{17}$$

In the same way, in this case where the expectation of $\mathcal{L}(X_1)$ is finite and thus $\int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x|>1\}} ds F(dx)$ exists, we can add

$$\int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x|>1\}} \left(\mu^X(ds, dx) - ds F(dx)\right) \tag{18}$$

to (10). Note that the sum of the big jumps, which is $\int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x|>1\}} \mu^X(ds, dx)$, always exists for every path. As a result of this we get (2) in the simpler representation

$$X_t = bt + \sqrt{c}W_t + \int_0^t \int_{\mathbb{R}} x\big(\mu^X(ds, dx) - dsF(dx)\big). \tag{19}$$

Of course the drift coefficient b in (19) is different from the drift coefficient b in the general representation (2). Actually b in (19) is nothing but the expectation $E[X_1]$ because the Brownian motion $(W_t)_{t\geq 0}$ and the pure jump integral process are both martingales with expectation zero and $E[X_t] = tE[X_1]$. From representation (19) it is immediately clear that (X_t) is a martingale if $b = E[X_1] = 0$, it is a submartingale if $b > 0$ and is a supermartingale if $b < 0$.

The case of finite expectation $E[X_1]$ under which we get (19) is of particular interest because all Lévy processes which we use in finance have finite first moments.

Whereas the frequency of the big jumps determines the existence of moments of the process, the *fine structure* of the paths of the process can be read off the frequency of the small jumps. We say the process has *finite activity* if almost all paths have only a finite number of jumps along any time interval of finite length. If almost all paths have infinitely many jumps along any time interval of finite length, we say the process has *infinite activity*.

Proposition 2 *Let $X = (X_t)_{t\geq 0}$ be a Lévy process with Lévy measure F.*

1. *X has finite activity if $F(\mathbb{R}) < \infty$.*
2. *X has infinite activity if $F(\mathbb{R}) = \infty$.*

Note that by definition a Lévy measure satisfies $\int_{\mathbb{R}} \mathbb{1}_{\{|x|>1\}} F(dx) < \infty$; therefore, the assumption $F(\mathbb{R}) < \infty$ or $F(\mathbb{R}) = \infty$ is equivalent to assuming finiteness or infiniteness of $\int_{\mathbb{R}} \mathbb{1}_{\{|x|\leq 1\}} F(dx) < \infty$. It is well known that the paths of Brownian motion have infinite variation. Consequently it follows from representation (2) or (19) that Lévy processes a priori have infinite variation paths as soon as $c > 0$. Whether the purely discontinuous component (Z_t) in (2) or the purely discontinuous integral process in (19) produces paths with finite or infinite variation again depends on the frequency of the small jumps.

Proposition 3 *Let $X = (X_t)_{t\geq 0}$ be a Lévy process with triplet (b, c, F).*

1. *Almost all paths of X have finite variation if $c = 0$ and $\int_{\{|x|\leq 1\}} |x| F(dx) < \infty$.*
2. *Almost all paths of X have infinite variation if $c \neq 0$ or $\int_{\{|x|\leq 1\}} |x| F(dx) = \infty$.*

For the proof see Theorem 21.9 in Sato (1999). The integrability of F in the sense that $\int_{\{|x|\leq 1\}} |x| F(dx) < \infty$ guarantees also that the sum of the

small jumps as given in (4) converges for (almost) every path. Therefore, in this case one can separate the integral in (10) or (19) and write, e.g., in (19)

$$\int_0^t \int_{\mathbb{R}} x\Big(\mu^X(ds,dx) - dsF(dx)\Big) = \int_0^t \int_{\mathbb{R}} x\mu^X(ds,dx) - t\int_{\mathbb{R}} xF(dx). \quad (20)$$

3 Financial Modeling

The classical model in finance for *stock prices* or *indices* which goes back to Samuelson (1965) and which became the basis for the Black–Scholes option pricing theory is the *geometric Brownian motion* given by the stochastic differential equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \quad (21)$$

This equation is solved by

$$S_t = S_0 \exp\big(\sigma W_t + (\mu - \sigma^2/2)t\big). \quad (22)$$

The exponent of this price process is a Lévy process as given in (2) with $b = \mu - \sigma^2/2$, $\sqrt{c} = \sigma$, $Z_t \equiv 0$ and no big jumps either. Log returns $\log S_{t+1} - \log S_t$ produced by this process along a time grid with span 1 are normally distributed variables $N(\mu - \sigma^2/2, \sigma^2)$ which are far from being realistic for most time series of financial data. Once one has identified a more realistic parametric distribution ν by fitting an empirical return distribution—several classes of candidate distributions will be presented in Section 4—the right starting point for a statistically more accurate model for prices is (22). Considering the Lévy process $X = (X_t)_{t\geq 0}$ such that $\mathcal{L}(X_1) = \nu$, the model

$$S_t = S_0 \exp(X_t), \quad (23)$$

which we call *exponential Lévy model*, produces along a time grid with span 1 log returns which are exactly equal to ν. This way one can implement an empirically derived (infinitely divisible) distribution into a model with exact returns. If one were to start instead with a stochastic differential equation, i.e., with the equivalent of (21), which is the equation

$$dS_t = S_{t-} dX_t, \quad (24)$$

one would get as the solution the stochastic exponential

$$S_t = S_0 \exp(X_t - ct/2) \prod_{s\leq t} (1 + \Delta X_s) \exp(-\Delta X_s). \quad (25)$$

The distribution of the log returns of this process is not known in general. As one can see directly from the term $(1+\Delta X_s)$ in (25), another drawback of this model is that it can produce negative stock prices as soon as the driving Lévy process X has negative jumps with absolute jump size larger than 1. The Lévy measures of all interesting classes of distributions which we shall consider in the next section have strictly positive densities on the whole negative half line and therefore the Lévy processes generated by these distributions have negative jumps of arbitrary size. For completeness we mention that the stochastic differential equation which describes the process (23) is

$$dS_t = S_{t-}\big(dX_t + (c/2)dt + \mathrm{e}^{\Delta X_t} - 1 - \Delta X_t\big). \tag{26}$$

For the pricing of derivatives it is interesting to characterize when the price process given by (23) is a *martingale* because pricing is done by taking expectations under a risk neutral or *martingale measure*. For $(S_t)_{t\geq 0}$ to be a martingale, first of all the expectation $E[S_t]$ has to be finite; therefore, candidates for the role of the driving process are Lévy processes X which have a finite first exponential moment

$$E[\exp(X_t)] < \infty. \tag{27}$$

Proposition 1 characterizes these processes in terms of their Lévy measure. At this point one should mention that the necessary assumption (27) a priori excludes stable processes as suitable driving processes for models in finance. Second, let X be given in the representation (19) (still assuming (27)) then $S_t = S_0 \exp(X_t)$ is a martingale if

$$b = -\frac{c}{2} - \int_{\mathbb{R}} \big(\mathrm{e}^x - 1 - x\big) F(dx). \tag{28}$$

This can be seen by applying Itô's formula to $S_t = S_0 \exp(X_t)$, where (28) guarantees that the drift component is 0. An alternative way to derive (28) is to verify that the process $(M_t)_{t\geq 0}$ given by

$$M_t = \frac{\exp(X_t)}{E[\exp(X_t)]} \tag{29}$$

is a martingale. Stationarity and independence of increments have to be used here. Equation (28) follows once one has verified (see (15)–(17)) that

$$E[\exp(X_t)] = \exp\left[t\left(b + \frac{c}{2} + \int_{\mathbb{R}} (\mathrm{e}^x - 1 - x) F(dx)\right)\right]. \tag{30}$$

The simplest models for fixed-income markets take the short rate r_t as the basic quantity and derive all other rates from r_t. More sophisticated approaches model simultaneously the whole term structure of rates for a continuum of maturities $[0, T^*]$ or as in the case of the LIBOR model the rates correspond-

ing to the maturities of a tenor structure $T_0 < T_1 < \cdots < T_N = T^*$. As a consequence these models are mathematically more demanding. A survey of interest rate modeling in the classic setting of diffusions is given by Björk (2008) in this volume. The interest rate theory for models driven by Lévy processes was developed in a series of papers by the author with various coauthors (Eberlein and Kluge (2006, 2007); Eberlein and Özkan (2003a, 2005); Eberlein and Raible (1999); Eberlein et al. (2005)).

Two basic approaches are the forward rate approach and the LIBOR approach. In the first case one assumes the dynamics of the *instantaneous forward rate* with maturity T, contracted at time t, $f(t,T)$ in the form

$$f(t,T) = f(0,T) + \int_0^t \alpha(s,T)ds - \int_0^t \sigma(s,T)dX_s \tag{31}$$

for any $T \in [0,T^*]$. The coefficients $\alpha(s,T)$ and $\sigma(s,T)$ can be deterministic or random. Starting with (31), one gets zero-coupon bond prices $B(t,T)$ in a form comparable to the stock price model (23), namely,

$$B(t,T) = B(0,T)\exp\left(\int_0^t (r(s) - A(s,T))ds + \int_0^t \Sigma(s,T)dX_s\right). \tag{32}$$

Here $r_s = r(s) = f(s,s)$ is the short rate and $A(s,T)$ and $\Sigma(s,T)$ are derived from $\alpha(s,T)$ and $\sigma(s,T)$ by integration.

In the *Lévy LIBOR market model* (Eberlein and Özkan (2005)) the forward LIBOR rates $L(t,T_j)$ for the time points $T_j (0 \le j \le N)$ of a tenor structure are chosen as the basic rates. As a result of a backward induction one gets for each j the rate in the following uniform form:

$$L(t,T_j) = L(0,T_j)\exp\left(\int_0^t \lambda(s,T_j)dX_s^{T_{j+1}}\right), \tag{33}$$

where $\lambda(s,T_j)$ is a volatility structure, $X^{T_{j+1}} = (X_t^{T_{j+1}})_{t\ge 0}$ is a process derived from an initial (time-homogeneous or time-inhomogeneous) Lévy process $X^{T_N} = (X_t^{T_N})_{t\ge 0}$ and (33) is considered under $\mathbb{P}^{T_{j+1}}$, the forward martingale measure which is derived during the backward induction. Closely related to the LIBOR model is the *forward process model*, where forward processes $F(t,T_j,T_{j+1}) = B(t,T_j)/B(t,T_{j+1})$ are chosen as the basic quantities and modeled in a form analogous to (33). An extension of the Lévy LIBOR approach to a multicurrency setting taking exchange rates into account was developed in Eberlein and Koval (2006). In all implementations of these models pure jump processes have been chosen as driving processes.

4 Examples of Lévy Processes with Jumps

4.1 Poisson and compound Poisson processes

The simplest Lévy measure one can consider is ε_1, a point mass in 1. Adding an intensity parameter $\lambda > 0$, one gets $F = \lambda\varepsilon_1$. Assuming $c = 0$, this Lévy measure generates a process $X = (X_t)_{t\geq 0}$ with jumps of size 1 which occur with an average rate of λ in a unit time interval. Otherwise the paths are constant. X is called a *Poisson process* with intensity λ. The drift parameter b in (17) is $E[X_1]$, which is λ. Therefore, the Fourier transform takes the form

$$E[\exp(\mathrm{i}uX_t)] = \exp\left[\lambda t(\mathrm{e}^{\mathrm{i}u} - 1)\right]. \tag{34}$$

Any variable X_t of the process has a Poisson distribution with parameter λt, i.e.,

$$P[X_t = k] = \exp(-\lambda t)\frac{(\lambda t)^k}{k!}.$$

One can show that the successive waiting times from one jump to the next are independent exponentially distributed random variables with parameter λ. Conversely, starting with a sequence $(\tau_i)_{i\geq 1}$ of independent exponentially distributed random variables with parameter λ and setting $T_n = \sum_{i=1}^{n} \tau_i$, the associated *counting process*

$$N_t = \sum_{n\geq 1} \mathbb{1}_{\{T_n \leq t\}} \tag{35}$$

is a Poisson process with intensity λ.

A natural extension of the Poisson process with jump height 1 is a process where the jump size is random. Let $(Y_i)_{i\leq 1}$ be a sequence of independent, identically distributed random variables with $\mathcal{L}(Y_1) = \nu$.

$$X_t = \sum_{i=1}^{N_t} Y_i, \tag{36}$$

where $(N_t)_{t\geq 0}$ is a Poisson process with intensity $\lambda > 0$ which is independent of $(Y_i)_{i\geq 1}$, defines a *compound Poisson process* $X = (X_t)_{t\geq 0}$ with intensity λ and jump size distribution ν. Its Fourier transform is given by

$$E[\exp(\mathrm{i}uX_t)] = \exp\left[\lambda t \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}ux} - 1\right)\nu(dx)\right]. \tag{37}$$

Consequently the Lévy measure is given by $F(A) = \lambda\nu(A)$ for measurable sets A in $\mathbb{R}$.

4.2 Lévy jump diffusion

A Lévy jump diffusion is a Lévy process where the jump component is given by a compound Poisson process. It can be represented in the form

$$X_t = bt + \sqrt{c}W_t + \sum_{i=1}^{N_t} Y_i, \tag{38}$$

where $b \in \mathbb{R}$, $c > 0$, $(W_t)_{t\geq 0}$ is a standard Brownian motion, $(N_t)_{t\geq 0}$ is a Poisson process with intensity $\lambda > 0$ and $(Y_i)_{i\geq 1})$ is a sequence of independent, identically distributed random variables which are independent of $(N_t)_{t\geq 0}$. For normally distributed random variables Y_i, Merton (1976) introduced the process (38) as a model for asset returns. Kou (2002) used double-exponentially distributed jump size variables Y_i. In principle any other distribution could be considered as well, but of course the question is if one can control explicitly the quantities one is interested in, for example, $\mathcal{L}(X_t)$.

4.3 Hyperbolic Lévy processes

Hyperbolic distributions which generate *hyperbolic Lévy processes* $X = (X_t : t \geq 0)$ – also called *hyperbolic Lévy motions* – constitute a four-parameter class of distributions. Their Lebesgue density is given by

$$d_H(x) = \frac{\sqrt{\alpha^2-\beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2-\beta^2})} \exp\big(-\alpha\sqrt{\delta^2+(x-\mu)^2}+\beta(x-\mu)\big). \tag{39}$$

Here K_ν denotes the modified Bessel function of the third kind with index ν. The four parameters of this distribution have the following meaning: $\alpha > 0$ determines the shape, β with $0 \leq |\beta| < \alpha$ the skewness, $\mu \in \mathbb{R}$ the location and $\delta > 0$ is a scaling parameter comparable to σ in the normal distribution. Taking the logarithm of d_H, one gets a hyperbola. This explains the name hyperbolic distribution. On the basis of an extensive empirical study of stock prices, hyperbolic Lévy processes were first used in finance in Eberlein and Keller (1995).

The Fourier transform ϕ_H of a hyperbolic distribution can be easily derived because of the exponential form of d_H in (39):

$$\phi_H(u) = \exp(\mathrm{i}u\mu)\left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+\mathrm{i}u)^2}\right)^{1/2} \frac{K_1(\delta\sqrt{\alpha^2-(\beta+\mathrm{i}u)^2})}{K_1(\delta\sqrt{\alpha^2-\beta^2})}. \tag{40}$$

Moments of all orders exist. In particular the expectation is given by

$$E[X_1] = \mu + \frac{\beta\delta}{\sqrt{\alpha^2-\beta^2}} \frac{K_2(\delta\sqrt{\alpha^2-\beta^2})}{K_1(\delta\sqrt{\alpha^2-\beta^2})}. \tag{41}$$

Analyzing ϕ_H in the form (17), one sees that $c = 0$. This means that hyperbolic Lévy motions are *purely discontinuous* processes. The Lévy measure F has an explicit *Lebesgue density* (see (46)).

4.4 Generalized hyperbolic Lévy processes

Hyperbolic distributions are a subclass of a more powerful five-parameter class, the *generalized hyperbolic distributions* (Barndorff-Nielsen (1978)). The additional class parameter $\lambda \in \mathbb{R}$ has the value 1 for hyperbolic distributions. The Lebesgue density for these distributions with parameters $\lambda, \alpha, \beta, \delta, \mu$ is

$$\begin{aligned} d_{GH}(x) = a(\lambda,\alpha,\beta,\delta)\big(\delta^2+(x-\mu)^2\big)^{(\lambda-\frac{1}{2})/2} \\ + K_{\lambda-\frac{1}{2}}\Big(\alpha\sqrt{\delta^2+(x-\mu)^2}\Big)\exp(\beta(x-\mu)), \end{aligned} \tag{42}$$

where the normalizing constant is given by

$$a(\lambda,\alpha,\beta,\delta) = \frac{(\alpha^2-\beta^2)^{\lambda/2}}{\sqrt{2\pi}\alpha^{\lambda-\frac{1}{2}}\delta^\lambda K_\lambda(\delta\sqrt{\alpha^2-\beta^2})}.$$

Other parameterizations are in use as well. Generalized hyperbolic distributions can be represented as normal mean-variance mixtures

$$d_{GH}(x;\lambda,\alpha,\beta,\delta,\mu) = \int_0^\infty d_{N(\mu+\beta y,y)}(x) d_{GIG}(y;\lambda,\delta,\sqrt{\alpha^2-\beta^2})dy, \tag{43}$$

where the mixing distribution is *generalized inverse Gaussian* with density

$$d_{GIG}(x;\lambda,\delta,\gamma) = \Big(\frac{\gamma}{\delta}\Big)^\lambda \frac{1}{2K_\lambda(\delta\gamma)} x^{\lambda-1}\exp\Big(-\frac{1}{2}\Big(\frac{\delta^2}{x}+\gamma^2 x\Big)\Big) \quad (x>0). \tag{44}$$

The exponential Lévy model with generalized hyperbolic Lévy motions as driving processes was introduced in Eberlein (2001) and Eberlein and Prause (2002).

The *moment generating function* $M_{GH}(u)$ exists for u with $|\beta+u| < \alpha$ and is given by

$$M_{GH}(u) = \exp(\mu u)\Big(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+u)^2}\Big)^{\lambda/2} \frac{K_\lambda(\delta\sqrt{\alpha^2-(\beta+u)^2})}{K_\lambda(\delta\sqrt{\alpha^2-\beta^2})}. \tag{45}$$

As a consequence, exponential moments (27) are finite. This fact is crucial for pricing of derivatives under martingale measures. The Fourier transform ϕ_{GH} is obtained from the relation $\phi_{GH}(u) = M_{GH}(\mathrm{i}u)$. Analyzing ϕ_{GH} in its form (17), we see again that $c = 0$, i.e., generalized hyperbolic Lévy motions are purely discontinuous processes. The Lévy measure F has a density given by

$$g_{GH}(x) \tag{46}$$
$$= \frac{\mathrm{e}^{\beta x}}{|x|}\Big(\int_0^\infty \frac{\exp\big(-\sqrt{2y+\alpha^2}|x|\big)}{\pi^2 y\big(J^2_{|\lambda|}(\delta\sqrt{2y}) + Y^2_{|\lambda|}(\delta\sqrt{2y})\big)} dy + \mathbb{1}_{\{\lambda\geq 0\}}\lambda \mathrm{e}^{-\alpha|x|}\Big).$$

Setting $\lambda = -\frac{1}{2}$ in (42) we get another interesting subclass, the *normal inverse Gaussian* (NIG) distributions, which were first used in finance in Barndorff-Nielsen (1998). Their Fourier transform is particularly simple since the Bessel function satisfies $K_{-1/2}(z) = K_{1/2}(z) = \sqrt{\pi/(2z)}\mathrm{e}^{-z}$. Therefore,

$$\phi_{NIG}(u) = \exp(\mathrm{i}u\mu)\exp\big(\delta\sqrt{\alpha^2-\beta^2}\big)\exp\big(-\delta\sqrt{\alpha^2-(\beta+\mathrm{i}u)^2}\big). \tag{47}$$

From the form (47) one sees immediately that NIG distributions are closed under convolution in the two parameters δ and μ, because taking a power t in (47), one gets the same form with parameters $t\delta$ and $t\mu$.

4.5 CGMY and variance gamma Lévy processes

Carr et al. (2002) introduced a class of infinitely divisible distributions—called CGMY—which extends the variance gamma model due to Madan and Seneta (1990) and Madan and Milne (1991). *CGMY Lévy processes* have purely discontinuous paths and the Lévy density is given by

$$g_{CGMY}(x) = \begin{cases} C\dfrac{\exp(-G|x|)}{|x|^{1+Y}} & x < 0, \\ C\dfrac{\exp(-Mx)}{x^{1+Y}} & x > 0. \end{cases} \tag{48}$$

The parameter space is C, G, $M > 0$ and $Y \in (-\infty, 2)$. The process has infinite activity if and only if $Y \in [0, 2)$ and the paths have infinite variation if and only if $Y \in [1, 2)$. For $Y = 0$ one gets the three-parameter *variance gamma distributions*. The latter are also a subclass of the generalized hyperbolic distributions (Eberlein and von Hammerstein (2004), Raible (2000)). For $Y < 0$ the Fourier transform of CGMY is given by

$$\phi_{CGMY}(u) = \exp\Big(C\Gamma(-Y)\big[(M-\mathrm{i}u)^Y - M^Y + (G+\mathrm{i}u)^Y - G^Y\big]\Big). \tag{49}$$

4.6 α-Stable Lévy processes

Stable distributions are a classical subject in probability. They constitute a four-parameter class of distributions with Fourier transform given by

$$\phi_{stab}(x) = \exp\left[\sigma^\alpha(-|\theta|^\alpha) + \mathrm{i}\theta\omega(\theta,\alpha,\beta)) + \mathrm{i}\mu\theta\right],$$

where

$$\omega(\theta,\alpha,\beta) = \begin{cases} \beta|\theta|^{\alpha-1}\tan\frac{\pi\alpha}{2} & \text{if } \alpha \neq 1, \\ -\beta\frac{2}{\pi}\ln|\theta| & \text{if } \alpha = 1. \end{cases} \tag{50}$$

The parameter space is $0 < \alpha \leq 2$, $\sigma \geq 0$, $-1 \leq \beta \leq 1$ and $\mu \in \mathbb{R}$. For $\alpha = 2$ one gets the Gaussian distribution with mean μ and variance $2\sigma^2$. For $\alpha < 2$ there is no Gaussian part, which means the paths of an α-stable Lévy motion are purely discontinuous in this case.

Explicit densities are known in three cases only: the Gaussian distribution ($\alpha = 2$, $\beta = 0$), the Cauchy distribution ($\alpha = 1$, $\beta = 0$) and the Lévy distribution ($\alpha = 1/2$, $\beta = 1$). Stable distributions have been used in risk management (Rachev and Mittnik (2000)), where the heavy tails are exploited. As pointed out earlier, their usefulness for modern financial theory in particular as a pricing model is limited for $\alpha \neq 2$ by the fact that the basic requirement (27) is not satisfied.

4.7 Meixner Lévy processes

The Fourier transform of Meixner distributions is given by

$$\phi_{Meix}(u) = \left(\frac{\cos(\beta/2)}{\cosh((\alpha u - \mathrm{i}\beta)/2)}\right)^{2\delta}$$

for parameters $\alpha > 0$, $|\beta| < \pi$, $\delta > 0$. The corresponding Lévy processes are purely discontinuous with Lévy density

$$g_{Meix}(x) = \delta\frac{\exp(\beta x/\alpha)}{x\sinh(\pi x/\alpha)}.$$

The process has paths of infinite variation. This process was introduced by Schoutens (2003) in the context of financial time series.

References

Bachelier, L. (1900): *Théorie de la spéculation*. PhD thesis, Annales Scientifiques de l'École Normale Superieure II **I–17**, 21–86. *English Translation in: Cootner, P. (Ed.) (1964): Random Character of Stock Market Prices*, 17–78. Massachusetts Institute of Technology, Cambridge.

Barndorff-Nielsen, O. E. (1978): Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics* **5**, 151–157.

Barndorff-Nielsen, O. E. (1998): Processes of normal inverse Gaussian type. *Finance and Stochastics*, **2**, 41–68.

Barndorff-Nielsen, O. E., Mikosch, T. and Resnick, S. I. (Eds.) (2001): *Lévy Processes. Theory and Applications*. Birkhäuser, Basel.

Bertoin, J. (1996): *Lévy processes*. Cambridge University Press, Cambridge.

Björk, T. (2008): An overview of interest rate theory. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 614–651. Springer, New York.

Black, F. and Scholes, M. (1973): The pricing of options and corporate liabilities. *Journal of Political Economy*, **81** 637–654.

Breiman, L. (1968): *Probability*. Addison-Wesley, Reading.

Carr, P., Geman, H., Madan, D. and Yor, M. (2002): The fine structure of asset returns: An empirical investigation. *Journal of Business* **75**, 305–332.

Cont, R. and Tankov, P. (2004): *Financial Modelling with Jump Processes*. Chapman & Hall/CRC, New York.

Eberlein, E. (2001): Application of generalized hyperbolic Lévy motions to finance. *In: Barndorff-Nielsen, O. E. et al. (Eds.): Lévy Processes. Theory and Applications*, 319–336. Birkhäuser, Basel.

Eberlein, E. and Keller, U. (1995): Hyperbolic distributions in finance. *Bernoulli*, **1**, 281–299.

Eberlein, E. and Kluge, W. (2006): Exact pricing formulae for caps and swaptions in a Lévy term structure model. *Journal of Computational Finance* **9**, 99–125.

Eberlein, E. and Kluge, W. (2007): Calibration of Lévy term structure models. *In: Fu, M., Jarrow, R. A., Yen, J.-Y. and Elliot, R. J. (Eds.): Advances in Mathematical Finance: In Honor of Dilip B. Madan*, 155–180. Birkhäuser, Basel.

Eberlein, E. and Koval, N. (2006): A cross-currency Lévy market model. *Quantitative Finance* **6**, 465–480.

Eberlein, E. and Özkan, F. (2003a): The defaultable Lévy term structure: Ratings and restructuring. *Mathematical Finance* **13**, 277–300.

Eberlein, E. and Özkan, F. (2003b): Time consistency of Lévy models. *Quantitative Finance* **3**, 40–50.

Eberlein, E. and Özkan, F. (2005): The Lévy LIBOR model. *Finance and Stochastics* **9**, 327–348.

Eberlein, E. and Prause, K. (2002): The generalized hyperbolic model: Financial derivatives and risk measures. *In: Mathematical Finance—Bachelier Congress, 2000 (Paris)*, 245–267. Springer, New York.

Eberlein, E. and Raible, S. (1999): Term structure models driven by general Lévy processes. *Mathematical Finance* **9**, 31–53.

Eberlein, E. and von Hammerstein, E. A. (2004): Generalized hyperbolic and inverse gaussian distributions: Limiting cases and approximation of processes. *In: Dalang, R. C., Dozzi, M., and Russo, F. (Eds.): Seminar on Stochastic Analysis, Random Fields and Applications IV, Progress in Probability* **58**, 221–264. Birkhäuser, Basel.

Eberlein, E., Jacod, J. and Raible, S. (2005): Lévy term structure models: No-arbitrage and completeness. *Finance and Stochastics* **9**, 67–88.

Fama, E. (1965): The behaviour of stock market prices. *Journal of Business* **38**, 34–105.

Jacod, J. and Shiryaev, A. N. (1987): *Limit Theorems for Stochastic Processes*. Springer, New York.

Kou, S. G. (2002): A jump diffusion model for option pricing. *Management Science* **48**, 1086–1101.

Madan, D. and Milne, F. (1991): Option pricing with V.G. martingale component. *Mathematical Finance* **1**, 39–55.

Madan, D. and Seneta, E. (1990): The variance gamma (V.G.) model for share market returns. *Journal of Business* **63**, 511–524.

Merton, R. C. (1976): Option pricing when underlying stock returns are discontinuous. *Journal Financ. Econ.* **3**, 125–144.

Protter, P. E. (2004): *Stochastic Integration and Differential Equations*. (2nd ed.). Volume 21 of *Applications of Mathematics*. Springer, New York.

Rachev, S. and Mittnik, S. (2000): *Stable Paretian Models in Finance*. Wiley, New York.

Raible, S. (2000): *Lévy processes in finance: Theory, numerics, and empirical facts*. PhD thesis, University of Freiburg.

Samorodnitsky, G. and Taqqu, M. (1999): *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*. Chapman & Hall/CRC, London.

Samuelson, P. (1965): Rational theory of warrant pricing. *Industrial Management Review* **6**:13–32.

Sato, K.-I. (1999): *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.

Schoutens, W. (2003): *Lévy Processes in Finance: Pricing Financial Derivatives*. Wiley, New York .

Lévy–Driven Continuous–Time ARMA Processes

Peter J. Brockwell

Abstract Gaussian ARMA processes with continuous time parameter, otherwise known as stationary continuous-time Gaussian processes with rational spectral density, have been of interest for many years. (See for example the papers of Doob (1944), Bartlett (1946), Phillips (1959), Durbin (1961), Dzhapararidze (1970,1971), Pham-Din-Tuan (1977) and the monograph of Arató (1982).) In the last twenty years there has been a resurgence of interest in continuous-time processes, partly as a result of the very successful application of stochastic differential equation models to problems in finance, exemplified by the derivation of the Black-Scholes option-pricing formula and its generalizations (Hull and White (1987)). Numerous examples of econometric applications of continuous-time models are contained in the book of Bergstrom (1990). Continuous-time models have also been utilized very successfully for the modelling of irregularly-spaced data (Jones (1981, 1985), Jones and Ackerson (1990)). Like their discrete-time counterparts, continuous-time ARMA processes constitute a very convenient parametric family of stationary processes exhibiting a wide range of autocorrelation functions which can be used to model the empirical autocorrelations observed in financial time series analysis. In financial applications it has been observed that jumps play an important role in the realistic modelling of asset prices and derived series such as volatility. This has led to an upsurge of interest in Lévy processes and their applications to financial modelling. In this article we discuss second-order Lévy-driven continuous-time ARMA models, their properties and some of their financial applications. Examples are the modelling of stochastic volatility in the class of models introduced by Barndorff-Nielsen and Shephard (2001) and the construction of a class of continuous-time GARCH models which generalize the COGARCH(1,1) process of Klüppelberg, Lindner and Maller (2004) and which exhibit properties analogous to those of the discrete-time GARCH(p, q) process.

Peter Brockwell
Department of Statistics, Colorado State University, Fort Collins, Colorado, 80523-1877, U.S.A., e-mail: pjbrock@stat.colostate.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_20,

1 Introduction

In financial econometrics, many discrete-time models (stochastic volatility, ARCH, GARCH and generalizations of these) are used to model the returns at regular intervals on stocks, currency investments and other assets. For example a GARCH process $(\xi_n)_{n\in\mathbb{N}}$ is frequently used to represent the increments, $\ln P_n - \ln P_{n-1}$, of the logarithms of the asset price P_n at times $1, 2, 3, \ldots$. These models capture many of the so-called *stylized features* of such data, e.g. tail heaviness, volatility clustering and dependence without correlation.

Various attempts have been made to capture the stylized features of financial time series using continuous-time models. The interest in continuous-time models stems from their use in modelling irregularly spaced data, their use in financial applications such as option-pricing and the current wide-spread availability of high-frequency data. In continuous-time it is natural to model the logarithm of the asset price itself, i.e. $G(t) = \ln P(t)$, rather than its increments as in discrete time.

One approach is via the stochastic volatility model of Barndorff-Nielsen and Shephard (2001) (see also Barndorff-Nielsen et al. (2002)), in which the volatility process V and the log asset price G satisfy the equations (apart from a deterministic rescaling of time),

$$dV(t) = -\lambda V(t)dt + dL(t), \tag{1.1}$$

$$dG(t) = (\gamma + \beta V(t))dt + \sqrt{V(t)}dW(t) + \rho dL(t), \tag{1.2}$$

where $\lambda > 0$, $L = (L(t))_{t\in\mathbb{R}_+}$ is a non-decreasing Lévy process and $W = (W(t))_{t\in\mathbb{R}_+}$ is standard Brownian motion independent of L. The volatility process V is taken to be a stationary solution of (1.1), in other words a *stationary Lévy-driven Ornstein-Uhlenbeck process* or a *continuous-time autoregression of order 1*. The background driving Lévy process L introduces the possibility of jumps in both the volatility and the log asset processes, a feature which is in accordance with empirical observations. It also allows for a rich class of marginal distributions, with possibly heavy tails. The autocorrelation function of the process V is $\rho(h) = \exp(-\lambda|h|)$. For modelling purposes this is quite restrictive, although the class of possible autocorrelations can be extended to a larger class of monotone functions if V is replaced by a superposition of such processes as in Barndorff-Nielsen (2001). However, as we shall see, a much wider class of not necessarily monotone autocorrelation functions for the volatility can be obtained by replacing the process V in (1.1) and (1.2) by a Lévy-driven continuous-time autoregressive moving average (CARMA) process as defined in Section 2. This class of processes constitutes a very flexible parametric family of stationary processes with a vast array of possible marginal distributions and autocorrelation functions.

Their role in continuous-time modelling is analogous to that of autoregressive moving average processes in discrete time. They belong to the more general class of Lévy-driven moving average processes considered by Fasen (2004).

A continuous-time analogue of the GARCH(1,1) process, denoted COGARCH(1,1), has recently been constructed and studied by Klüppelberg et al. (2004). Their construction is based on an explicit representation of the discrete-time GARCH(1,1) volatility process which they use in order to obtain a continuous-time analogue. Since no such representation exists for higher-order discrete-time GARCH processes, a different approach is needed to construct higher-order models in continuous time. The Lévy-driven CARMA process plays a key role in this construction.

The present paper deals with *second-order* Lévy-driven continuous-time ARMA (denoted CARMA) processes, since for most financial applications processes with finite second moments are generally considered adequate. (Analogous processes without the second-order assumption are considered in Brockwell (2001).) In Section 2 we review the definition and properties, deriving the kernel and autocovariance functions, specifying the joint characteristic functions and discussing the issue of causality. Under the assumption of distinct autoregressive roots, some particularly tractable representations of the kernel, the autocovariance function and the process itself are derived. The question of recovering the driving process from a realization of the process on a (continuous) interval $[0, T]$ is also considered.

Section 3 considers connections between continuous-time and discrete-time ARMA processes.

In Section 4 we indicate the applications of CARMA processes to the modelling of stochastic volatility in the Barndorff-Nielsen-Shephard stochastic volatility model and in Section 5 their role in the construction of COGARCH models of order higher than (1,1).

Section 6 deals briefly with the well-established methods of inference for Gaussian CARMA processes and the far less developed question of inference for more general Lévy-driven processes.

Before proceeding further we need a few essential facts regarding Lévy processes. (For a detailed account of the pertinent properties of Lévy processes see Protter (2004) and for further properties see the books of Applebaum (2004), Bertoin (1996) and Sato (1999).) Suppose we are given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\leq t\leq\infty}, P)$, where $\mathcal{F}_0$ contains all the P-null sets of $\mathcal{F}$ and $(\mathcal{F}_t)$ is right-continuous.

Definition 1 (Lévy Process)
An adapted process $\{L(t), t \geq 0\}$ is said to be a Lévy process if

(i) $L(0) = 0$ a.s.,
(ii) $L(t) - L(s)$ is independent of $\mathcal{F}_s$, $0 \leq s < t < \infty$,
(iii) $L(t) - L(s)$ has the same distribution as $L(t - s)$ and
(iv) $L(t)$ is continuous in probability.

Every Lévy process has a unique modification which is càdlàg (right continuous with left limits) and which is also a Lévy process. We shall therefore assume that our Lévy process has these properties. The characteristic function of $L(t)$, $\phi_t(\theta) := E(\exp(i\theta L(t)))$, has the Lévy-Khinchin representation,

$$\phi_t(\theta) = \exp(t\xi(\theta)),\ \theta \in \mathbb{R}, \tag{1.3}$$

where

$$\xi(\theta) = \mathrm{i}\theta m - \frac{1}{2}\theta^2 s^2 + \int_{\mathbb{R}_0} (e^{\mathrm{i}\theta x} - 1 - \mathrm{i}x\theta I_{\{|x|<1\}})\nu(dx), \tag{1.4}$$

for some $m \in \mathbb{R}$, $s \geq 0$, and measure ν on the Borel subsets of $\mathbb{R}_0 = \mathbb{R}\backslash\{0\}$. ν is known as the *Lévy measure* of the process L and satisfies the condition $\int_{\mathbb{R}_0} \min(1, |u|^2)\nu(du) < \infty$. If ν is the zero measure then $\{L(t)\}$ is Brownian motion with $E(L(t)) = mt$ and $\mathrm{Var}(L(t)) = s^2 t$. If $s^2 = 0$, $\nu(\mathbb{R}_0) < \infty$ and $a = \int_{\mathbb{R}_0} uI_{\{|u|<1\}}(u)\nu(du)$, then $\{L(t)\}$ is a compound Poisson process with jump-rate $\nu(\mathbb{R}_0)$ and jump-size distribution $\nu/\nu(\mathbb{R}_0)$. A wealth of distributions for $L(t)$ is attainable by suitable choice of the measure ν. See for example Barndorff-Nielsen and Shephard (2001). For the second-order Lévy processes (with which we are concerned in this paper), $E(L(1))^2 < \infty$. To avoid problems of parameter identifiability we shall assume throughout that L is scaled so that $\mathrm{Var}(L(1)) = 1$. Then $\mathrm{Var}(L(t)) = t$ for all $t \geq 0$ and there exists a real constant μ such that $EL(t) = \mu t$ for all $t \geq 0$. We shall then refer to the process L as a *standardized second-order Lévy process*, written henceforth as SSLP.

2 Second–Order Lévy–Driven CARMA Processes

A second-order Lévy-driven continuous-time ARMA(p, q) process, where p and q are non-negative integers such that $q < p$, is defined (see Brockwell (2001)) via the state-space representation of the formal equation,

$$a(D)Y(t) = \sigma b(D)DL(t),\ t \geq 0, \tag{2.1}$$

where σ is a strictly positive scale parameter, D denotes differentiation with respect to t, $\{L(t)\}$ is an SSLP,

$$a(z) := z^p + a_1 z^{p-1} + \cdots + a_p,$$

$$b(z) := b_0 + b_1 z + \cdots + b_{p-1} z^{p-1},$$

and the coefficients b_j satisfy $b_q = 1$ and $b_j = 0$ for $q < j < p$. The behaviour of the process is determined by the process L and the coefficients $\{a_j, 1 \leq j \leq p;\ b_j, 0 \leq j < q;\ \sigma\}$. In view of the scale parameter, σ, on the right-hand side

of (2.1), there is clearly no loss of generality in assuming that $\mathrm{Var}(L(1)) = 1$, i.e. that L is an SSLP as defined at the end of Section 1. To avoid trivial and easily eliminated complications we shall assume that $a(z)$ and $b(z)$ have no common factors. The state-space representation consists of the *observation* and *state* equations,

$$Y(t) = \sigma \mathbf{b}' \mathbf{X}(t), \tag{2.2}$$

and

$$d\mathbf{X}(t) - A\mathbf{X}(t)dt = \mathbf{e}\, dL(t), \tag{2.3}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_p & -a_{p-1} & -a_{p-2} & \cdots & -a_1 \end{bmatrix}, \ \mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-2} \\ b_{p-1} \end{bmatrix}.$$

(If $p = 1$, A is defined to be $-a_1$.) In the special case when $\{L(t)\}$ is standard Brownian motion, (2.3) is an Ito equation with solution $\{\mathbf{X}(t), t \geq 0\}$ satisfying

$$\mathbf{X}(t) = e^{At}\mathbf{X}(0) + \int_0^t e^{A(t-u)}\mathbf{e}\, dL(u), \tag{2.4}$$

where the integral is defined as the L^2 limit of approximating Riemann-Stieltjes sums S_n corresponding to the partition of the interval $[0, t]$ by the points $\{k/2^n, k \in \mathbf{Z}, 0 \leq k < 2^n t\}$ and $\{t\}$. If L is any second-order Lévy process the integral is defined in the same way. The continuous differentiability of the integrand in (2.4) implies that the sequence $\{S_n\}$ converges geometrically rapidly in L^2 and hence almost surely to the same limit. In fact the integral in (2.4) is a special case (with deterministic and continuously differentiable integrand) of integration with respect to a semimartingale as discussed in the book of Protter (2004). From (2.4) we can also write

$$\mathbf{X}(t) = e^{A(t-s)}\mathbf{X}(s) + \int_s^t e^{A(t-u)}\mathbf{e}\, dL(u), \text{ for all } t > s \geq 0, \tag{2.5}$$

which clearly shows (by the independence of increments of $\{L(t)\}$) that $\{\mathbf{X}(t)\}$ is Markov. The following propositions give necessary and sufficient conditions for stationarity of $\{\mathbf{X}(t)\}$.

Proposition 1 *If* $\mathbf{X}(0)$ *is independent of* $\{L(t), t \geq 0\}$ *and* $E(L(1)^2) < \infty$, *then* $\{\mathbf{X}(t)\}$ *is weakly stationary if and only if the eigenvalues of the matrix*

A all have strictly negative real parts and $\mathbf{X}(0)$ *has the mean and covariance matrix of* $\int_0^\infty e^{Au}\mathbf{e}\, dL(u)$, *i.e.* $-A^{-1}\mathbf{e}\mu$ *and* $\int_0^\infty e^{Ay}\mathbf{e}\,\mathbf{e}'e^{A'y}dy$ *respectively.*

Proof The eigenvalues of A must have negative real parts for the sum of the covariance matrices of the terms on the right of (2.4) to be bounded in t. If this condition is satisfied then $\{\mathbf{X}(t)\}$ converges in distribution as $t \to \infty$ to a random variable with the distribution of $\int_0^\infty e^{Au}\mathbf{e}\, dL(u)$. Hence, for weak stationarity, $\mathbf{X}(0)$ must have the mean and covariance matrix of $\int_0^\infty e^{Au}\mathbf{e}\, dL(u)$. Conversely if the eigenvalues of A all have negative real parts and if $\mathbf{X}(0)$ has the mean and covariance matrix of $\int_0^\infty e^{Au}\mathbf{e}\, dL(u)$, then a simple calculation using (2.4) shows that $\{\mathbf{X}(t)\}$ is weakly stationary.

Proposition 2 *If* $\mathbf{X}(0)$ *is independent of* $\{L(t), t \geq 0\}$ *and* $E(L(1)^2) < \infty$, *then* $\{\mathbf{X}(t)\}$ *is strictly stationary if and only if the eigenvalues of the matrix A all have strictly negative real parts and* $\mathbf{X}(0)$ *has the distribution of* $\int_0^\infty e^{Au}\mathbf{e}\, dL(u)$.

Proof Necessity follows from Proposition 1. If the conditions are satisfied then strict stationarity follows from the fact that $\{\mathbf{X}(t)\}$ is a Markov process whose initial distribution is the same as its limit distribution.

Remark 1 It is convenient to extend the state process $\{\mathbf{X}(t), t \geq 0\}$ to a process with index set $(-\infty, \infty)$. To this end we introduce a second Lévy process $\{M(t), 0 \leq t < \infty\}$, independent of L and with the same distribution, and then define the following extension of L:

$$L^*(t) = L(t)I_{[0,\infty)}(t) - M(-t-)I_{(-\infty,0]}(t),\ -\infty < t < \infty.$$

Then, provided the eigenvalues of A all have negative real parts, the process $\{\mathbf{X}(t)\}$ defined by

$$\mathbf{X}(t) = \int_{-\infty}^{t} e^{A(t-u)}\mathbf{e}\, dL^*(u), \tag{2.6}$$

is a strictly stationary process satisfying (2.5) (with L replaced by L^*) for all $t > s$ and $s \in (-\infty, \infty)$. Henceforth we shall refer to L^* as the background SSLP and denote it for simplicity by L rather than L^*.

Remark 2 It is easy to check that the eigenvalues of the matrix A, which we shall denote by $\lambda_1, \ldots, \lambda_p$, are the same as the zeroes of the autoregressive polynomial $a(z)$. The corresponding right eigenvectors are

$$[1\ \lambda_j\ \lambda_j^2\ \ldots\ \lambda_j^{p-1}]',\ j = 1, \ldots, p,$$

We are now in a position to define the CARMA process $\{Y(t), -\infty < t < \infty\}$ under the condition that

$$\mathcal{R}e(\lambda_j) < 0,\ j = 1, \ldots, p. \tag{2.7}$$

Definition 2 (Causal CARMA Process)
If the zeroes $\lambda_1, \ldots, \lambda_p$ of the autoregressive polynomial $a(z)$ satisfy (2.7), then the CARMA(p, q) process driven by the SSLP $\{L(t), -\infty < t < \infty\}$ with coefficients $\{a_j, 1 \le j \le p;\ b_j, 0 \le j < q;\ \sigma\}$ is the strictly stationary process,

$$Y(t) = \sigma \mathbf{b}' \mathbf{X}(t),$$

where

$$\mathbf{X}(t) = \int_{-\infty}^{t} e^{A(t-u)} \mathbf{e}\ dL(u),$$

i.e.

$$Y(t) = \sigma \int_{-\infty}^{t} \mathbf{b}' e^{A(t-u)} \mathbf{e}\ dL(u). \tag{2.8}$$

Remark 3 (Causality and Non-causality)
Under Condition (2.7) we see from (2.8) that $\{Y(t)\}$ is a *causal* function of $\{L(t)\}$, since it has the form

$$Y(t) = \int_{-\infty}^{\infty} g(t-u)\ dL(u), \tag{2.9}$$

where

$$g(t) = \begin{cases} \sigma \mathbf{b}' e^{At} \mathbf{e}, & \text{if t>0,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

The function g is referred to as the *kernel* of the CARMA process $\{Y(t)\}$. Under the condition (2.7), the function g defined by (2.10) can be written as

$$g(t) = \frac{\sigma}{2\pi} \int_{-\infty}^{\infty} e^{\mathrm{i}t\lambda} \frac{b(\mathrm{i}\lambda)}{a(\mathrm{i}\lambda)}\ d\lambda. \tag{2.11}$$

(To establish (2.11) when the eigenvalues $\lambda_1, \ldots, \lambda_p$ are distinct, we use the explicit expressions for the eigenvectors of A to replace e^{At} in (2.10) by its spectral representation. The same expression is obtained when the right side of (2.11) is evaluated by contour integration. When there are multiple eigenvalues, the result is obtained by separating the eigenvalues slightly and taking the limit as the repeated eigenvalues converge to their common value.) It is of interest to observe that the representation (2.9) and (2.11) of $\{Y(t)\}$ defines a strictly stationary process even under conditions less restrictive than (2.7), namely

$$\mathcal{R}e(\lambda_j) \neq 0,\ j = 1, \ldots, p.$$

Thus (2.9) and (2.11) provide a more general definition of CARMA process than Definition 2 above. However if any of the zeroes of $a(z)$ has real part greater than 0, the representation (2.9) of $\{Y(t)\}$ in terms of $\{L(t)\}$ will no

longer be causal as is the case when (2.7) is satisfied. This distinction between causal and non-causal CARMA processes is analogous to the classification of discrete-time ARMA processes as causal or otherwise, depending on whether or not the zeroes of the autoregressive polynomial lie outside the unit circle (see e.g. Brockwell and Davis (1991)). From now on **we shall restrict attention to causal CARMA processes**, i.e. we shall assume that (2.7) holds, so that the general expression (2.11) for the kernel g can also be written in the form (2.10). However both forms of the kernel will prove to be useful.

Remark 4 (Second-order Properties)
From the representation (2.8) of a causal CARMA process driven by the SSLP L with $EL(1) = \mu$, we immediately find that

$$EY(t) = -\sigma \mathbf{b}' A^{-1} \mathbf{e} \mu = \sigma \mu b_0 / a_p$$

and

$$\gamma(h) := \operatorname{cov}(Y(t+h), Y(t)) = \sigma^2 \mathbf{b}' \, e^{A|h|} \Sigma \, \mathbf{b}, \tag{2.12}$$

where

$$\Sigma = \int_0^\infty e^{Ay} \mathbf{e} \, \mathbf{e}' e^{A'y} dy.$$

From the representation (2.9) of $Y(t)$ we see that γ can also be expressed as

$$\gamma(h) = \operatorname{cov}(Y(t+h), Y(t)) = \int_{-\infty}^\infty \widetilde{g}(h-u) g(u) du,$$

where $\widetilde{g}(x) = g(-x)$ and g is defined in (2.11). Using the convolution theorem for Fourier transforms, we find that

$$\int_{-\infty}^\infty e^{-\mathrm{i}\omega h} \gamma(h) dh = \sigma^2 \left| \frac{b(\mathrm{i}\omega)}{a(\mathrm{i}\omega)} \right|^2,$$

showing that the spectral density of the process Y is

$$f(\omega) = \frac{\sigma^2}{2\pi} \left| \frac{b(\mathrm{i}\omega)}{a(\mathrm{i}\omega)} \right|^2 \tag{2.13}$$

and the autocovariance function is

$$\gamma(h) = \frac{\sigma^2}{2\pi} \int_{-\infty}^\infty e^{\mathrm{i}\omega h} \left| \frac{b(\mathrm{i}\omega)}{a(\mathrm{i}\omega)} \right|^2 d\omega. \tag{2.14}$$

The spectral density (2.13) is clearly a rational function of the frequency ω. The family of Gaussian CARMA processes is in fact identical to the class of stationary Gaussian processes with rational spectral density.

Remark 5 (Distinct Autoregressive Zeroes, the Canonical State Representation and Simulation of Y)
When the zeroes $\lambda_1, \ldots, \lambda_p$ of $a(z)$ are distinct and satisfy the causality condition (2.7), the expression for the kernel g takes an especially simple form. Expanding the integrand in (2.11) in partial fractions and integrating each term gives the simple expression,

$$g(h) = \sigma \sum_{r=1}^{p} \frac{b(\lambda_r)}{a'(\lambda_r)} e^{\lambda_r h} I_{[0,\infty)}(h). \tag{2.15}$$

Applying the same argument to (2.14) gives a corresponding expression for the autocovariance function, i.e.

$$\gamma(h) = \mathrm{cov}(Y_{t+h}, Y_t) \;=\; \sigma^2 \sum_{j=1}^{p} \frac{b(\lambda_j) b(-\lambda_j)}{a'(\lambda_j) a(-\lambda_j)} e^{\lambda_j |h|}. \tag{2.16}$$

When the autoregressive roots are distinct we obtain a very useful representation of the CARMA(p, q) process Y from (2.15). Defining

$$\alpha_r = \sigma \frac{b(\lambda_r)}{a'(\lambda_r)}, \; r = 1, \ldots, p, \tag{2.17}$$

we can write

$$Y(t) = \sum_{r=1}^{p} Y_r(t), \tag{2.18}$$

where

$$Y_r(t) = \int_{-\infty}^{t} \alpha_r e^{\lambda_r (t-u)} dL(u). \tag{2.19}$$

This expression shows that the component processes Y_r satisfy the simple equations,

$$Y_r(t) = Y_r(s) e^{\lambda_r (t-s)} + \int_{s}^{t} \alpha_r e^{\lambda_r (t-u)} dL(u), \; t \geq s, \; r = 1, \ldots, p. \tag{2.20}$$

Taking $s = 0$ and using Lemma 2.1 of Eberlein and Raible (1999), we find that

$$Y_r(t) = Y_r(0) e^{\lambda_r t} + \alpha_r L(t) + \int_{0}^{t} \alpha_r \lambda_r e^{\lambda_r (t-u)} L(u) du, \; t \geq 0, \tag{2.21}$$

where the last integral is a Riemann integral and the equality holds for all finite $t \geq 0$ with probability 1. Defining

$$\mathbf{Y}(t) := [Y_1(t), \ldots, Y_p(t)]', \tag{2.22}$$

we obtain from (2.6), (2.15) and (2.19),

$$\mathbf{Y}(t) = \sigma B R^{-1} \mathbf{X}(t), \tag{2.23}$$

where $B = \text{diag}[b(\lambda_i)]_{i=1}^p$ and $R = [\lambda_j^{i-1}]_{i,j=1}^p$. The initial values $Y_r(0)$ in (2.21) can therefore be obtained from those of the components of the state vector $\mathbf{X}(0)$. The process $\mathbf{Y}$ provides us with an alternative *canonical* state representation of $Y(t),\ t \geq 0$, namely

$$Y(t) = [1, \ldots, 1]\mathbf{Y}(t) \tag{2.24}$$

where $\mathbf{Y}$ is the solution of

$$d\mathbf{Y}(t) = \text{diag}[\lambda_i]_{i=1}^p \mathbf{Y} dt + \sigma B R^{-1} \mathbf{e}\ dL. \tag{2.25}$$

with $\mathbf{Y}(0) = \sigma B R^{-1} \mathbf{X}(0)$.

Notice that the canonical representation of the process Y reduces the problem of simulating CARMA(p, q) processes with distinct autoregressive roots to the much simpler problem of simulating the (possibly complex-valued) component CAR(1) processes (2.19) and adding them together.

Example 1 (The CAR(1) Process)
The CAR(1) (or stationary Ornstein-Uhlenbeck) process satisfies (2.1) with $b(z) = 1$ and $a(z) = z - \lambda$ where $\lambda < 0$. From (2.15) and (2.16) we immediately find that $g(h) = e^{\lambda h} I_{[0,\infty)}(h)$ and $\gamma(h) = \sigma^2 e^{\lambda|h|}/(2|\lambda|)$. In this case the 1×1 matrices B and R are both equal to 1 so the (1-dimensional) state vectors $\mathbf{X}$ and $\mathbf{Y}$ are identical and the state-space representation given by (2.2) and (2.3) is already in canonical form. Equations (2.18) and (2.19) reduce to

$$Y(t) = Y_1(t)$$

and

$$Y_1(t) = \sigma \int_{-\infty}^{t} e^{\lambda(t-u)} dL(u)$$

respectively (since $\lambda_1 = \lambda$ and $\alpha_1 = \sigma$).

Example 2 (The CARMA(2,1) Process)
In this case $b(z) = b_0 + z$, $a(z) = (z - \lambda_1)(z - \lambda_2)$ and the real parts of λ_1 and λ_2 are both negative. Assuming that $\lambda_1 \neq \lambda_2$, we find from (2.15) that

$$g(h) = (\alpha_1 e^{\lambda_1 h} + \alpha_2 e^{\lambda_2 h}) I_{[0,\infty)}(h)$$

where $\alpha_r = \sigma(b_0 + \lambda_r)/(\lambda_r - \lambda_{3-r}),\ r = 1, 2$. An analogous expression for $\gamma(h)$ can be found from (2.16). From (2.23) the canonical state vector is

$$\mathbf{Y}(t) = \begin{bmatrix} Y_1(t) \\ Y_2(t) \end{bmatrix} = \frac{\sigma}{\lambda_1 - \lambda_2} \begin{bmatrix} \lambda_2(b_0 + \lambda_1) & -(b_0 + \lambda_1) \\ -\lambda_1(b_0 + \lambda_2) & b_0 + \lambda_2 \end{bmatrix} \mathbf{X}(t)$$

and the canonical representation of Y is, from (2.18) and (2.19),

$$Y(t) = Y_1(t) + Y_2(t)$$

where

$$Y_r(t) = \int_{-\infty}^{t} \alpha_r e^{\lambda_r(t-u)} dL(u), \ r = 1, 2,$$

and $\alpha_r = \sigma(b_0 + \lambda_r)/(\lambda_r - \lambda_{3-r}), \ r = 1, 2.$

Remark 6 (The Joint Distributions)
Since the study of Lévy-driven CARMA processes is largely motivated by the need to model processes with non-Gaussian joint distributions, it is important to go beyond a second-order characterization of these processes. From Proposition 2 we already know that the marginal distribution of $Y(t)$ is that of $\int_0^\infty g(u)dL(u)$, where g is given by (2.11) or, if the autoregressive roots are distinct and the causality conditions (2.7) are satisfied, by (2.15). Using the expression (1.3) for the characteristic function of $L(t)$, we find that the cumulant generating function of $Y(t)$ is

$$\log E(\exp(\mathrm{i}\theta Y(t))) = \int_0^\infty \xi(\theta g(u))du, \tag{2.26}$$

showing that the distribution of $Y(t)$, like that of $L(t)$, is infinitely divisible. In the special case of the CAR(1) process the distribution of $Y(t)$ is also self-decomposable (see Barndorff-Nielsen and Shephard (2001), Theorem 2.1, and the accompanying references). More generally it can be shown (see Brockwell (2001)) that the cumulant generating function of $Y(t_1), Y(t_2), \ldots, Y(t_n)$, $(t_1 < t_2 < \cdots < t_n)$ is

$$\log E[\exp(\mathrm{i}\theta_1 Y(t_1) + \cdots + i\theta_n Y(t_n))] =$$

$$\int_0^\infty \xi\left(\sum_{i=1}^n \theta_i g(t_i + u)\right) du + \int_0^{t_1} \xi\left(\sum_{i=1}^n \theta_i g(t_i - u)\right) du + \tag{2.27}$$

$$\int_{t_1}^{t_2} \xi\left(\sum_{i=2}^n \theta_i g(t_i - u)\right) du + \cdots + \int_{t_{n-1}}^{t_n} \xi\left(\theta_n g(t_n - u)\right) du.$$

If $\{L(t)\}$ is a compound Poisson process with finite jump-rate λ and bilateral exponential jump-size distribution with probability density $f(x) = \frac{1}{2}\beta e^{-\beta|x|}$, then the corresponding CAR(1) process (see Example 1) has marginal cumulant generating function,

$$\kappa(\theta) = \int_0^\infty \xi(\theta e^{-cu})du,$$

where $\xi(\theta) = \lambda\theta^2/(\beta^2 + \theta^2)$. Straightforward evaluation of the integral gives

$$\kappa(\theta) = -\frac{\lambda}{2c} \ln\left(1 + \frac{\theta^2}{\beta^2}\right),$$

showing that $Y(t)$ has a symmetrized gamma distribution, or more specifically that $Y(t)$ is distributed as the difference between two independent gamma distributed random variables with exponent $\lambda/(2c)$ and scale parameter β. In particular, if $\lambda = 2c$, the marginal distribution is bilateral exponential. For more examples see Barndorff-Nielsen and Shephard (2001).

Remark 7 (Recovering the driving noise process)
For statistical modelling, one needs to know or to postulate an appropriate family of models for the driving Lévy process L. It would be useful therefore to recover the realized driving process, for given or estimated values of $\{a_j, 1 \le j \le p; b_j, 0 \le j < q; \sigma\}$, from a realization of Y on some finite interval $[0, T]$. This requires knowledge of the initial state vector $\mathbf{X}(0)$ in general, but if this is available (as for example when a CARMA$(p, 0)$ process is observed continuously on $[0, T]$), or if we are willing to assume a plausible value for $\mathbf{X}(0)$, then an argument due to Pham-Din-Tuan (1977) can be used to recover $\{L(t), 0 \le t \le T\}$. We shall assume in this Remark that the polynomial b (as well as the polynomial a) has all its zeroes in the left half-plane. This assumption is analogous to that of invertibility in discrete time. Since the covariance structure of our Lévy-driven process is exactly the same (except for slight notational changes) as that of Pham-Din-Tuan's Gaussian CARMA process and since his result holds for Gaussian CARMA processes with arbitrary mean (obtained by adding a constant to the zero-mean process) his L^2-based spectral argument can be applied directly to the Lévy-driven CARMA process to give, for $t \ge 0$,

$$L(t) = \sigma^{-1}\left[Y^{(p-q-1)}(t) - Y^{(p-q-1)}(0)\right] - \int_0^t \left[\sum_{j=1}^{q} b_{q-j} X^{(p-j)}(s) - \sum_{j=1}^{p} a_j X^{(p-j)}(s)\right] ds, \tag{2.28}$$

where $Y^{(p-q-1)}$ denotes the derivative of order $p-q-1$ of the CARMA process Y and $X^{(0)}, \ldots, X^{(p-1)}$ are the components of the state process $\mathbf{X}$ (the component $X^{(j)}$ being the j^{th} derivative of $X^{(0)}$). $\mathbf{X}(t)$ can be expressed in terms of Y and $\mathbf{X}(0)$ by noting that (2.2) characterizes $X^{(0)}$as a CARMA$(q, 0)$ process driven by the process $\{\sigma^{-1}\int_0^t Y(s)ds\}$. Making use of this observation, introducing the $q \times 1$ state vector $\mathbf{X}_q(t) := [X^{(0)}(t), \ldots, X^{(q-1)}(t)]'$ and proceeding exactly as we did in solving the CARMA equations in Section 2, we find that, for $q \ge 1$,

$$\mathbf{X}_q(t) = \mathbf{X}_q(0)e^{Bt} + \sigma^{-1}\int_0^t e^{B(t-u)}\mathbf{e}_q Y(u)du, \tag{2.29}$$

where

$$B = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -b_0 & -b_1 & -b_2 & \cdots & -b_{q-1} \end{bmatrix} \text{ and } \mathbf{e}_q = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

with $B := -b_0$ if $q = 1$, while for $q = 0$,

$$X^{(0)}(t) = \sigma^{-1}Y(t). \tag{2.30}$$

The remaining derivatives of $X^{(0)}$ up to order $p-1$ can be determined from (2.29) or (2.30), completing the determination of the state vector $\mathbf{X}(t)$. Having recovered $\mathbf{X}(t)$, the SSLP is found from (2.28).

To illustrate the use of (2.28) and (2.29) or (2.30), we consider the CAR(1) process of Example 1. In this case $a(z) = z - \lambda$, $b(z) = 1$ and the (one-dimensional) state vector is, from (2.30), $X(t) = \sigma^{-1}Y(t)$. Substituting into (2.28) gives

$$L(t) = \sigma^{-1}\left[Y(t) - Y(0) - \lambda\int_0^t Y(s)ds\right]. \tag{2.31}$$

It is easy to check directly that if Y is a Lévy-driven CARMA(1,0) process with parameters $a_1(= -\lambda)$ and σ and if L is the Lévy process defined by (2.31), then

$$Y(t) = Y(0)e^{\lambda t} + \sigma\int_0^t e^{\lambda(t-u)}dL(u), \tag{2.32}$$

since the last integral can be rewritten, by Lemma 2.1 of Eberlein and Raible (1999), as $\sigma L(t) + \sigma\int_0^t \lambda e^{\lambda(t-u)}L(u)du$. Making this replacement and substituting from (2.31) for L, we see that the right-hand side of (2.32) reduces to $Y(t)$.

In the case when the autoregressive roots are distinct, we can use the transformation (2.23) to recover the canonical state process $\mathbf{Y}$ defined by (2.19) and (2.22) from $\mathbf{X}$. Then applying the argument of Pham-Din-Tuan to the component processes Y_r we obtain p (equivalent) representations of $L(t)$, namely

$$L(t) = \alpha_r^{-1}\left[Y_r(t) - Y_r(0) - \lambda_r\int_0^t Y_r(s)ds\right],\ r = 1, \ldots, p. \tag{2.33}$$

Although Pham-Din-Tuan's result was derived with real-valued processes in mind, it is easy to check directly, as in the CARMA(1,0) case, that if Y is a

Lévy-driven CARMA(p,q) process with parameters $\{a_j, 1 \le j \le p;\ b_j, 0 \le j < q;\ \sigma\}$ and L is the Lévy process satisfying the equations (2.33) with possibly complex-valued Y_r and λ_r, then

$$Y_r(t) = Y_r(0)e^{\lambda_r t} + \int_0^t \alpha_r e^{\lambda_r(t-u)} dL(u),\ t \ge 0,\ r = 1, \ldots, p,$$

and these equations imply, with (2.23), that the state process $\mathbf{X}$ satisfies

$$\mathbf{X}(t) = e^{At}\mathbf{X}(0) + \int_0^t e^{A(t-u)}\mathbf{e} dL(u),\ t \ge 0,$$

showing that $Y = \sigma \mathbf{b}'\mathbf{X}$ is indeed the CARMA(p,q) process with parameters $\{a_j, 1 \le j \le p;\ b_j, 0 \le j < q;\ \sigma\}$ driven by L. Thus we have arrived at p very simple (equivalent) representations of the driving SSLP, any of which can be computed from the realization of Y, the value of $\mathbf{X}(0)$ and the parameters of the CARMA process. Of course for calculations it is simplest to choose a value of r in (2.34) for which λ_r is real (if such an r exists).

3 Connections with Discrete–Time ARMA Processes

The discrete-time ARMA(p,q) process $\{Y_n\}$ with autoregressive coefficients $\phi_1, \ldots, \phi_p$, moving average coefficients $\theta_1, \ldots, \theta_q$, and white noise variance σ_d^2, is defined to be a (weakly) stationary solution of the p^{th} order linear difference equations,

$$\phi(B)Y_n = \theta(B)Z_n,\ n = 0, \pm 1, \pm 2, \ldots, \tag{3.1}$$

where B is the backward shift operator ($BY_n = Y_{n-1}$ and $BZ_n = Z_{n-1}$ for all n), $\{Z_n\}$ is a sequence of uncorrelated random variables with mean zero and variance σ_d^2 (abbreviated to $\{Z_n\} \sim \text{WN}(0, \sigma_d^2)$) and

$$\phi(z) := 1 - \phi_1 z - \cdots - \phi_p z^p,$$

$$\theta(z) := 1 + \theta_1 z + \cdots + \theta_q z^q,$$

with $\theta_q \ne 0$ and $\phi_p \ne 0$. We define $\phi(z) := 1$ if $p = 0$ and $\theta(z) := 1$ if $q = 0$. We shall assume that the polynomials $\phi(z)$ and $\theta(z)$ have no common zeroes and that $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ is non-zero for all complex z such that $|z| \le 1$. This last condition guarantees the existence of a unique stationary solution of (3.1) which is also causal, i.e. is expressible in the form $Y_n = \sum_{j=0}^{\infty} \psi_j Z_{n-j}$ for some absolutely summable sequence $\{\psi_j\}$. It is evident from this representation that the mean of the ARMA process defined by (3.1) is zero. The process $\{Y_n\}$ is said to be an ARMA(p,q) process with

mean μ if $\{Y_n - \mu\}$ is an ARMA(p, q) process. A more restrictive definition of ARMA process imposes the further requirement that the random variables Z_n be independent and identically distributed, in which case we write $\{Z_n\} \sim \text{IID}(0, \sigma_d^2)$. The process $\{Y_n\}$ is then strictly (as well as weakly) stationary and we shall refer to $\{Y_n\}$ as a *strict ARMA process*. If we impose the further constraint that each Z_n is Gaussian, then we write $\{Z_n\} \sim \text{IIDN}(0, \sigma_d^2)$ and refer to $\{Y_n\}$ as a *Gaussian ARMA process*.

As one might expect, there are many structural similarities between ARMA and CARMA processes. In the case when the reciprocals $\xi_1, \ldots, \xi_p$ of the zeroes of the polynomial $\phi(z)$ are distinct and $q < p$, there is an analogue of (2.16) for the autocovariance function of the ARMA process, namely

$$\gamma_d(h) = -\sigma_d^2 \sum_{j=1}^{p} \frac{\xi_j^{|h|+1} \theta(\xi_j)\theta(\xi_j^{-1})}{\phi(\xi_j)\phi'(\xi_j^{-1})}, \ h = 0, \pm 1, \pm 2, \ldots. \tag{3.2}$$

There is also a corresponding *canonical representation* analogous to that in Remark 5 of Section 2. It takes the form (cf. (2.18) and (2.19)),

$$Y_n = \sum_{r=1}^{p} Y_{r,n}, \tag{3.3}$$

and

$$Y_{r,n} = \sum_{k=-\infty}^{n} \beta_r \xi_r^{n-k} Z_k, \ r = 1, \ldots, p \tag{3.4}$$

where ξ_r, $r = 1, \ldots, p$, are the reciprocals of the(distinct) zeroes of $\phi(z)$, and

$$\beta_r = -\xi_r \frac{\theta(\xi_r^{-1})}{\phi'(\xi_r^{-1})}, \ r = 1, \ldots, p. \tag{3.5}$$

From (3.4) we also obtain the relations (cf. (2.20)),

$$Y_{r,n} = \xi_r Y_{r,n-1} + \beta_r Z_n, \ n = 0, \pm 1, \ldots; \ r = 1, \ldots, p. \tag{3.6}$$

Remark 8 When $q < p$ and the autoregressive roots are distinct, the equations (2.19) and (3.6) show that both the CARMA and ARMA processes can be represented as a sum of autoregressive processes of order 1. Note however that in both cases the component processes are not independent and are in general complex valued.

Example 3 (The AR(1) Process)
The defining equation (3.1) with $\phi(z) = 1 - \xi z$ and $\theta(z) = 1$ is clearly already in canonical form and, since $\beta_1 = 1$, equations (3.3) and (3.4) take the form

$$Y_n = Y_{1,n}$$

where

$$Y_{1,n} = \sum_{k=-\infty}^{n} \xi^{n-k} Z_k. \tag{3.7}$$

Example 4 (The ARMA(2,1) Process)
In this case $\phi(z) = (1-\xi_1 z)(1-\xi_2 z)$, where we assume that $|\xi_1| < 1$, $|\xi_2| < 1$ and $\xi_1 \neq \xi_2$. The moving average polynomial is $\theta(z) = 1 + \theta_1 z$ and the white noise variance is σ_d^2. From (3.5) we find that

$$\beta_r = \frac{\xi_r + \theta_1}{\xi_r - \xi_{3-r}}, \ r = 1, 2. \tag{3.8}$$

The canonical representation of the ARMA(2,1) process is thus

$$Y_n = Y_{1,n} + Y_{2,n},$$

where

$$Y_{r,n} = \beta_r \sum_{k=-\infty}^{n} \xi_r^{n-k} Z_k, \ r = 1, 2, \tag{3.9}$$

with β_r, $r = 1, 2$, as defined in (3.7).

If Y is a Gaussian CARMA process defined as in Section 2 with standard Brownian motion as the driving process, then it is well-known (see e.g. Doob (1944), Phillips (1959), Brockwell (1995)) that the sampled process $(Y(n\delta))_{n\in\mathbf{Z}}$ with fixed $\delta > 0$ is a (strict) Gaussian ARMA(r, s) process with $0 \leq s < r \leq p$ and spectral density

$$f_\delta(\omega) = \sum_{k=-\infty}^{\infty} \delta^{-1} f_Y(\delta^{-1}(\omega + 2k\pi)), \ -\pi \leq \omega \leq \pi, \tag{3.10}$$

where $f_Y(\omega)$, $-\infty < \omega < \infty$, is the spectral density of the original CARMA process.

If L is non-Gaussian, the sampled process will have the same spectral density and autocovariance function as the process obtained by sampling a Gaussian CARMA process with the same parameters, driven by Brownian motion with the same mean and variance as L. Consequently from a second-order point of view the two sampled processes will be the same. However, except in the case of the CAR(1) process, the sampled process will not generally be a strict ARMA process.

If Y is the CAR(1) process in Example 1, the sampled process is the strict AR(1) process satisfying

$$Y(n\delta) = e^{\lambda\delta} Y((n-1)\delta) + Z_n, \ n = 0, \pm 1, \ldots, \tag{3.11}$$

where

$$Z_n = \sigma \int_{(n-1)\delta}^{n\delta} e^{\lambda(n\delta-u)} dL(u). \tag{3.12}$$

The noise sequence $\{Z(n)\}$ is i.i.d. and $Z(n)$ has the infinitely divisible distribution with log characteristic function $\int_0^\delta \xi(\sigma\theta e^{\lambda u})du$, where $\xi(\theta)$ is the log characteristic function of $L(1)$ as in (1.3). For the CARMA(p,q) process with $p > 1$ the situation is more complicated. If the autoregressive roots $\lambda_1, \ldots, \lambda_p$, are all distinct, then from (2.18) and (2.19) the sampled process $\{Y(n\delta)\}$ is the sum of the strict AR(1) component processes $\{Y_r(n\delta)\}$, $r = 1, \ldots, p$, satisfying

$$Y_r(n\delta) = e^{\lambda_r \delta} Y_r((n-1)\delta) + Z_r(n), \ n = 0, \pm 1, \ldots,$$

where

$$Z_r(n) = \alpha_r \int_{(n-1)\delta}^{n\delta} e^{\lambda_r(n\delta-u)} dL(u),$$

and α_r is given by (2.17).

The following question is important if we estimate parameters of a CARMA process by fitting a discrete-time ARMA(p,q) process with $q < p$ to regularly spaced data and then attempt to find the parameters of a CARMA process whose values at the observation times have the same distribution as the values of the fitted ARMA process at those times. The critical question here is whether or not such a CARMA process exists.

If a given Gaussian ARMA(p,q) process with $q < p$ is distributed as the observations at integer times of *some* Gaussian CARMA process it is said to be *embeddable*. Embeddability depends on the polynomials $\phi(z)$ and $\theta(z)$. Many, but not all, Gaussian ARMA processes are embeddable. For example the ARMA(1,0) process (3.1) with $\phi(z) = 1 - \phi_1 z$ and white-noise variance σ_d^2 can be embedded, if $0 < \phi < 1$, in the Gaussian CAR(1) process defined by (2.1) with $a(z) = z - \log(\phi_1)$, $b(z) = 1$ and $\sigma^2 = -2\log(\phi_1)\sigma_d^2/(1-\phi_1^2)$ and, if $-1 < \phi_1 < 0$, it can be embedded in a CARMA(2,1) process (see Chan and Tong (1987)). However Gaussian ARMA processes for which $\theta(z) = 0$ has a root on the unit circle are not embeddable in *any* CARMA process (see Brockwell and Brockwell (1999)). The class of non-embeddable Gaussian ARMA processes also includes ARMA(2,1) processes with autocovariance functions of the form $\gamma(h) = C_1\xi_1^{|h|} + C_2\xi_2^{|h|}$, where ξ_1 and ξ_2 are distinct values in $(0,1)$ and $C_1\log(\xi_1) + C_2\log(\xi_2) > 0$. Such ARMA processes exist since there are infinitely many values of C_1 and C_2 satisfying the latter condition for which γ is a non-negative-definite function on the integers.

The problem of finding a CARMA process whose *autocovariance function* at integer lags matches that of a given non-Gaussian ARMA process is clearly equivalent to the problem of embedding a Gaussian ARMA process as described above.

However the determination of a Lévy-driven CARMA process (if there is one) whose sampled process has the same *joint distributions* as a given non-Gaussian ARMA process is more difficult. For example, from (3.11) and (3.12) we see that in order to embed a discrete-time AR(1) in a CAR(1) process, the driving noise sequence $\{Z_n\}$ of the AR(1) process must be i.i.d. with an infinitely divisible distribution, and the coefficient ϕ in the autoregressive polynomial $(1-\phi z)$ must be positive. Given such a process, with coefficient $\phi \in (0,1)$ and white-noise characteristic function $\exp(\psi(\theta))$, it is embeddable in a CAR(1) process (which must have autoregressive polynomial $a(z) = z - \lambda$, where $\lambda = \log(\phi)$) if and only if there exists a characteristic function $\exp(\rho(\theta))$ such that

$$\int_0^1 \rho(\theta e^{\lambda u})du = \psi(\theta), \text{ for all } \theta \in \mathbb{R}, \tag{3.13}$$

and then $\exp(\rho(\theta)t)$ is the characteristic function of $\sigma L(t)$ for the CAR(1) process in which the AR(1) process can be embedded. It is easy to check that if $\psi(\theta) = -\sigma_d^2\theta^2/2$, i.e. if Z_n is normally distributed with mean zero and variance σ_d^2, then (3.13) is satisfied if $\rho(\theta) = -\lambda\sigma_d^2\theta^2/(1-e^{2\lambda})$, i.e. if $\sigma L(1)$ is normally distributed with mean zero and variance $2\lambda\sigma_d^2/(1-e^{2\lambda})$. (More generally if Z_n is symmetric α-stable with $\psi(\theta) = -c|\theta|^\alpha$, $c > 0$, $\alpha \in (0,2]$, (3.13) is satisfied if $\rho(\theta) = -\alpha c\lambda|\theta|^\alpha/(1-e^{2\lambda})$, i.e. if $\sigma L(1)$ also has a symmetric α-stable distribution. If $\alpha \in (0,2)$ the processes do not have finite variance but the embedding is still valid.)

4 An Application to Stochastic Volatility Modelling

In the stochastic volatility model (1.1) and (1.2) of Barndorff-Nielsen and Shephard, the volatility process V is a CAR(1) (or stationary Ornstein-Uhlenbeck) process driven by a non-decreasing Lévy process L. With this model the authors were able to derive explicit expressions for quantities of fundamental interest, such as the integrated volatility. Since the process V can be written,

$$V(t) = \int_{-\infty}^{t} e^{-\lambda(t-u)}dL(u),$$

and since both the kernel, $g(u) = e^{-\lambda u}I_{(0,\infty)}(u)$, and the increments of the driving Lévy process are non-negative, the volatility is non-negative as required. A limitation of the use of the Ornstein-Uhlenbeck process however (and of linear combinations with non-negative coefficients of independent Ornstein-Uhlenbeck processes) is the constraint that the autocorrelations $\rho(h)$, $h \geq 0$, are necessarily non-increasing in h.

Much of the analysis of Barndorff-Nielsen and Shephard can however be carried out after replacing the Ornstein-Uhlenbeck process by a CARMA pro-

cess with non-negative kernel driven by a non-decreasing Lévy process. This has the advantage of allowing the representation of volatility processes with a larger range of autocorrelation functions than is possible in the Ornstein-Uhlenbeck framework. For example, the CARMA(3,2) process with

$$a(z) = (z + 0.1)(z + 0.5 - \mathrm{i}\pi/2)(z + 0.5 - \mathrm{i}\pi/2) \text{ and } b(z) = 2.792 + 5z + z^2$$

has non-negative kernel and autocovariance functions,

$$g(t) = 0.8762e^{-0.1t} + \left(0.1238 \cos \frac{\pi t}{2} + 2.5780 \sin \frac{\pi t}{2}\right) e^{-0.5t},\ t \geq 0,$$

and

$$\gamma(h) = 5.1161e^{-0.1h} + \left(4.3860 \cos \frac{\pi h}{2} + 1.4066 \sin \frac{\pi h}{2}\right) e^{-0.5h},\ h \geq 0,$$

respectively, both of which exhibit damped oscillatory behaviour.

There is of course a constraint imposed upon the allowable CARMA processes for stochastic volatility modelling by the requirement that the kernel g be non-negative. Conditions on the coefficients which guarantee non-negativity of the kernel have been considered by Brockwell and Davis (2001) and Todorov and Tauchen (2004) for the CARMA(2,1) process with real autoregressive roots and, more generally by Tsai and Chan (2004). In his analysis of the German Mark/US Dollar exchange rate series from 1986 through 1999, Todorov (2005) finds that a good fit to the autocorrelation function of the realized volatility is provided by a CARMA(2,1) model with two real autoregessive roots.

A class of long-memory Lévy-driven CARMA processes was introduced by Brockwell (2004) and Brockwell and Marquardt (2005) by replacing the kernel g in (2.9) by the kernel,

$$g_d(t) = \int_{-\infty}^{\infty} e^{\mathrm{i}t\lambda} (\mathrm{i}\lambda)^{-d} \frac{b(\mathrm{i}\lambda)}{a(\mathrm{i}\lambda)} d\lambda,$$

with $0 < d < 0.5$. The resulting processes, which exhibit hyperbolic rather than geometric decay in their autocorrelation functions, must however be driven by Lévy processes with zero mean, and such Lévy processes cannot be non-decreasing. Long-memory Lévy-driven CARMA processes cannot therefore be used directly for the modelling of stochastic volatility. They can however be used for the modelling of mean-corrected log volatility in order to account for the frequently observed long memory in such series.

5 Continuous–Time GARCH Processes

A continuous-time analog of the GARCH(1,1) process, denoted COGARCH(1,1), has recently been constructed and studied by Klüppelberg et al. (2004). Their construction uses an explicit representation of the discrete-time GARCH(1,1) process to obtain a continuous-time analog. Since no such representation exists for higher-order discrete-time GARCH processes, a different approach is needed to construct higher-order continuous-time analogs. For a detailed discussion of continuous-time GARCH processes see the article of Lindner (2008) in the present volume.

Let $(\varepsilon_n)_{n\in N_0}$ be an iid sequence of random variables. For any non-negative integers p and q, the discrete-time GARCH(p,q) process $(\xi_n)_{n\in\mathbb{N}_0}$ is defined by the equations,

$$\begin{aligned} \xi_n &= \sigma_n\varepsilon_n, \\ \sigma_n^2 &= \alpha_0 + \alpha_1\xi_{n-1}^2 + \ldots + \alpha_p\xi_{n-p}^2 + \beta_1\sigma_{n-1}^2 + \ldots + \beta_q\sigma_{n-q}^2, \end{aligned} \tag{5.1}$$

where $s := \max(p, q)$, the initial values $\sigma_0^2, \ldots, \sigma_{s-1}^2$ are assumed to be iid and independent of the iid sequence $(\varepsilon_n)_{n\geq s}$, and $\xi_n = G_{n+1} - G_n$ represents the increment at time n of the log asset price process $(G_n)_{n\in\mathbb{N}_0}$. In continuous-time it is more convenient to define the GARCH process as a model for $(G_t)_{t\geq 0}$ rather than for its increments as in discrete-time.

Equation (5.1) shows that the volatility process $(V_n := \sigma_n^2)_{n\in\mathbb{N}_0}$ can be viewed as a "self-exciting" ARMA$(q, p-1)$ process driven by the noise sequence $(V_{n-1}\varepsilon_{n-1}^2)_{n\in\mathbb{N}}$. This observation suggests defining a continuous time GARCH model of order (p, q) for the log asset price process $(G_t)_{t\geq 0}$ by

$$dG_t = \sqrt{V_t}\, dL_t, \quad t > 0, \quad G_0 = 0,$$

where $(V_t)_{t\geq 0}$ is a left-continuous non-negative CARMA$(q, p-1)$ process driven by a suitable replacement for the discrete time driving noise sequence $(V_{n-1}\varepsilon_{n-1}^2)_{n\in\mathbb{N}}$. By choosing the driving process to be

$$R_t = \int_0^t V_s d[L, L]_s^{(d)}, \quad \text{i.e.} \quad dR_t = V_t\, d[L, L]_t^{(d)}.$$

where $[L, L]^{(d)}$ is the *discrete part of the quadratic covariation* of the Lévy process L, we obtain the COGARCH(p, q) process, which has properties analogous to those of the discrete-time GARCH process and which includes the COGARCH(1,1) process of Klüppelberg et al.(2004) as a special case. The precise definition is as follows.

Definition 3 (COGARCH(p, q) process)
If p and q are integers such that $q \geq p \geq 1$, $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_p \in \mathbb{R}$, $\beta_1, \ldots, \beta_q \in \mathbb{R}$, $\alpha_p \neq 0$, $\beta_q \neq 0$, and $\alpha_{p+1} = \ldots = \alpha_q = 0$, we define the $(q \times q)$–matrix B and the vectors $\mathbf{a}$ and $\mathbf{e}$ by

$$B = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\beta_q & -\beta_{q-1} & -\beta_{q-2} & \dots & -\beta_1 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{q-1} \\ \alpha_q \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

with $B := -\beta_1$ if $q = 1$. Then if $L = (L_t)_{t\geq 0}$ is a Lévy process with non-trivial Lévy measure, we define the (left-continuous) *volatility process* $V = (V_t)_{t\geq 0}$ with parameters B, $\mathbf{a}$, α_0 and driving Lévy process L by

$$V_t = \alpha_0 + \mathbf{a}'\mathbf{Y}_{t-}, \quad t > 0, \quad V_0 = \alpha_0 + \mathbf{a}'\mathbf{Y}_0,$$

where the *state process* $\mathbf{Y} = (\mathbf{Y}_t)_{t\geq 0}$ is the unique càdlàg solution of the stochastic differential equation

$$d\mathbf{Y}_t = B\mathbf{Y}_{t-}\, dt + \mathbf{e}(\alpha_0 + \mathbf{a}'\mathbf{Y}_{t-})\, d[L, L]_t^{(d)}, \quad t > 0,$$

with initial value $\mathbf{Y}_0$, independent of the driving Lévy process $(L_t)_{t\geq 0}$. If the process $(V_t)_{t\geq 0}$ is strictly stationary and non-negative almost surely, we say that $G = (G_t)_{t\geq 0}$, given by

$$dG_t = \sqrt{V_t}\, dL_t, \quad t > 0, \quad G_0 = 0,$$

is a *COGARCH*(p, q) *process* with parameters B, $\mathbf{a}$, α_0 and driving Lévy process L .

Conditions for the existence of a non-negative stationary solution of the equations for V and the properties of the resulting volatility and COGARCH(p, q) processes, including conditions for the existence of moments of order k, are studied in the paper of Brockwell et al. (2006). In particular it is shown under mild conditions that the process of increments

$$G_t^{(r)} := G_{t+r} - G_t,$$

for any fixed $r > 0$, has the characteristic GARCH properties,

$$EG_t^{(r)} = 0, \ \operatorname{cov}(G_{t+h}^{(r)}, G_t^{(r)}) = 0 \ h \geq r,$$

while the squared increment process $G^{(r)2}$ has a non-zero autocovariance function, expressible in terms of the defining parameters of the process. The autocovariance function of the stationary volatility process, if it exists, is that of a CARMA process, just as the discrete-time GARCH volatility process has the autocovariance function of an ARMA process.

6 Inference for CARMA Processes

Given observations of a CARMA(p, q) process at times $0 \le t_1 < t_2 < \ldots < t_N$, there is an extensive literature on maximum *Gaussian* likelihood estimation of the parameters. This literature however does not address the question of identifying and estimating parameters for the driving process when it is not Gaussian. In the general case we can write, from (2.2) and (2.5),

$$Y(t_i) = \sigma \mathbf{b}' \mathbf{X}(t_i), \ i = 1, \ldots, N, \tag{6.1}$$

where

$$\mathbf{X}(t_i) = e^{A(t_i - t_{i-1})} \mathbf{X}(t_{i-1}) + \int_{t_{i-1}}^{t_i} e^{A(t_i - u)} \mathbf{e} \, dL(u), \ i = 2, \ldots, N, \tag{6.2}$$

and $\mathbf{X}(t_1)$ has the distribution of $\int_0^\infty e^{Au} \mathbf{e} dL(u)$. The *observation equations* (6.1) and *state equations* (6.2) are in precisely the form required for application of the discrete-time Kalman recursions (see e.g. Brockwell and Davis (1991)) in order to compute numerically the best one-step linear predictors of $Y_2, \ldots, Y_N$, and hence the Gaussian likelihood of the observations in terms of the coefficients $\{a_j, 1 \le j \le p; \ b_j, 0 \le j < q; \ \sigma\}$. Jones (1981) used this representation, together with numerical maximization of the calculated Gaussian likelihood, to compute maximum Gaussian likelihood estimates of the parameters for time series with irregularly spaced data. A similar approach was used in a more general setting by Bergstrom (1985). If the observations are uniformly spaced an alternative approach due to Phillips (1959) is to fit a discrete-time ARMA model to the observations and then to determine a Gaussian CARMA process in which the discrete-time process can be embedded. (Recalling the results of Section 3 however, it may be the case that there is no CARMA process in which the fitted ARMA process can be embedded.)

For a CAR(p) process observed continuously on the time interval $[0, T]$, Hyndman (1993) derived continuous-time analogues of the discrete-time Yule-Walker equations for estimating the coefficients. For a Gaussian CARMA process observed continuously on $[0, T]$, the exact likelihood function was determined by Pham-Din-Tuan (1977) who also gave a computational algorithm for computing approximate maximum likelihood estimators of the parameters which are asymptotically normal and efficient. The determination of the exact likelihood, conditional on the initial state vector $\mathbf{X}(0)$, can also be carried out for non-linear Gaussian CAR(p) processes and maximum conditional likelihood estimators expressed in terms of stochastic integrals (see Brockwell et al. (2006), where this method of estimation is applied to threshold CAR processes observed at closely spaced times, using sums to approximate the stochastic integrals involved.)

For Lévy-driven CARMA processes, estimation procedures which take into account the generally non-Gaussian nature of L are less well-developed. One

approach is to estimate the parameters $\{a_j, 1 \leq j \leq p;\ b_j, 0 \leq j < q;\ \sigma\}$ by maximizing the Gaussian likelihood of the observations using (6.1) and (6.2). If the process is observed continuously on $[0, T]$, these estimates and the results of Remark 2.8 can be used to recover, for any observed or assumed $\mathbf{X}(0)$, a realization of L on $[0, T]$. The increments of this realization can then be examined and a driving Lévy process chosen whose increments are compatible with the increments of the recovered realization of L. If the CARMA process is observed at closely-spaced discrete time points then a discretized version of this procedure can be used. Inference for stationary Ornstein-Uhlenbeck processes with non-decreasing driving Lévy process has been investigated by Jongbloed et al. (2006) and Brockwell et al. (2007).

Acknowledgement I am indebted to the National Science Foundation and Deutsche Forschungsgemeinschaft for support of this work under grants DMS-0308109, DMS-0744058 and SFB 386. I am also indebted to Alexander Lindner and Yu Yang for valuable comments on an early version of the manuscript.

References

Applebaum, D. (2004): *Lévy Processes and Stochastic Calculus* Cambridge University Press, Cambridge.

Arató, M. (1982): *Linear Stochastic Systems with Constant Coefficients* Springer Lecture Notes in Control and Information Systems **45**, Springer-Verlag, Berlin.

Barndorff-Nielsen, O.E. (2001): Superposition of Ornstein-Uhlenbeck type processes. *Theory Probab. Appl.* **45**, 175–194.

Barndorff-Nielsen, O.E. and Shephard, N. (2001): Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics (with discussion). *J. Roy. Statist. Soc. Ser. B* **63**, 167–241.

Barndorff-Nielsen, O.E., Nicolato E. and Shephard, N. (2002): Some recent developments in stochastic volatility modelling. *Quantitative Finance* **2**, 11–23.

Bartlett, M.S. (1946): On the theoretical specification and sampling properties of autocorrelated time series. *J. Royal Statistical Soc. (Supplement)* **7**, 27–41.

Bergstrom, A.R. (1985): The estimation of parameters in non-stationary higher-order continuous-time dynamic models. *Econometric Theory* **1**, 369–385.

Bergstrom, A.R. (1990): *Continuous Time Econometric Modelling* Oxford University Press, Oxford.

Bertoin, J. (1996): *Lévy Processes* Cambridge University Press, Cambridge.

Brockwell, A.E. and Brockwell, P.J. (1998): A class of non-embeddable ARMA processes. *J. Time Series Analysis* **20**, 483–486.

Brockwell, P.J. (1995): A note on the embedding of discrete-time ARMA processes. *J. Time Series Analysis* **16**, 451–460.

Brockwell, P.J. (2000): Continuous–time ARMA processes. *In: C.R. Rao and D.N. Shanbhag (Eds.): Stochastic processes: theory and methods, Handbook of Statist.* **19**, 249–276. North–Holland, Amsterdam.

Brockwell, P.J. (2001): Lévy-driven CARMA processes. *Ann. Inst. Stat. Mat.* **53**, 113–124.

Brockwell, P.J. (2004): Representations of continuous-time ARMA processes. *J. Appl. Probab.* **41A**, 375–382.

Brockwell, P.J. and Davis, R.A. (1991): *Time Series: Theory and Methods* 2nd edition. Springer, New York.

Brockwell, P.J. and Marquardt, T. (2005): Fractionally integrated continuous-time ARMA processes. *Statistica Sinica* **15**, 477–494.

Brockwell, P.J., Chadraa, E., and Lindner, A. (2006): Continuous-time GARCH processes. *Annals Appl. Prob.* **16**, 790–826.

Brockwell, P.J, Davis, R.A. and Yang, Y. (2007): Continuous-time autoregression. *Statistica Sinica* **17**, 63–80.

Brockwell, P.J, Davis, R.A. and Yang, Y. (2007): Inference for non-negative Lévy-driven Ornstein-Uhlenbeck processes. *J. Appl. Prob.* **44**, 977–989.

Chan, K.S. and Tong, H. (1987): A Note on embedding a discrete parameter ARMA model in a continuous parameter ARMA model. *J. Time Ser. Anal.* **8**, 277–281.

Doob, J.L. (1944): The elementary Gaussian processes. *Ann. Math. Statist.* **25**, 229–282.

Durbin, J. (1961): Efficient fitting of linear models for continuous stationary time series from discrete data. *Bull. Int. Statist. Inst.* **38**, 273–281.

Dzhaparidze, K.O. (1970): On the estimation of the spectral parameters of a stationary Gaussian process with rational spectral density. *Th. Prob. Appl.* **15**, 531–538.

Dzhaparidze, K.O. (1971): On methods for obtaining asymptotically efficient spectral parameter estimates for a stationary Gaussian process with rational spectral density. *Th. Prob. Appl.* **16**, 550–554.

Eberlein, E. and Raible, S. (1999): Term structure models driven by general Lévy processes. *Mathematical Finance* **9**, 31–53.

Fasen, V. (2004): *Lévy Driven MA Processes with Applications in Finance.* Ph.D. thesis, Technical University of Munich.

Hull, J. and White, A. (1987): The pricing of assets on options with stochastic volatilities. *J. of Finance* **42**, 281–300.

Hyndman, R.J. (1993): Yule-Walker estimates for continuous-time autoregressive models. *J. Time Series Analysis* **14**, 281–296.

Jones, R.H. (1981): Fitting a continuous time autoregression to discrete data. *In: Findley, D.F. (Ed.): Applied Time Series Analysis II*, 651–682. Academic Press, New York.

Jones, R.H. (1985): Time series analysis with unequally spaced data. *In: Hannan, E.J., Krishnaiah, P.R. and Rao, M.M. (Eds.): Time Series in the Time Domain, Handbook of Statistics* **5**, 157–178. North Holland, Amsterdam.

Jones, R.H. and Ackerson, L.M. (1990): Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**, 721–732.

Jongbloed, G., van der Meulen, F.H. and van der Vaart, A.W. (2005): Non-parametric inference for Lévy-driven Ornstein-Uhlenbeck processes. *Bernoulli* **11**, 759–791.

Klüppelberg, C., Lindner, A. and Maller, R. (2004): A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour. *J. Appl. Probab.* **41**, 601–622.

Lindner, A. (2008): Continuous Time Approximations to GARCH and Stochastic Volatility Models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 481–496. Springer, New York.

Pham-Din-Tuan (1977): Estimation of parameters of a continuous-time Gaussian stationary process with rational spectral density function. *Biometrika* **64**, 385–399.

Phillips, A.W. (1959): The estimation of parameters in systems of stochastic differential equations. *Biometrika* **46**, 67–76.

Protter, P.E. (2004): *Stochastic Integration and Differential Equations.* 2nd edition. Springer, New York.

Sato, K. (1999): *Lévy Processes and Infinitely Divisible Distributions* Cambridge University Press, Cambridge.

Todorov, V. (2007): Econometric analysis of jump-driven stochastic volatility models. *www.kellogg.northwestern.edu/faculty/todorov/htm/papers/jdsv.pdf*

Todorov, V. and Tauchen, G. (2006): Simulation methods for Lévy-driven CARMA stochastic volatility models. *J. Business and Economic Statistics* **24**, 455–469.

Tsai, H. and Chan, K.S. (2005): A note on non-negative continuous-time processes. *J. Roy. Statist. Soc. Ser. B* **67**, 589–597.

Continuous Time Approximations to GARCH and Stochastic Volatility Models

Alexander M. Lindner

Abstract We collect some continuous time GARCH models and report on how they approximate discrete time GARCH processes. Similarly, certain continuous time volatility models are viewed as approximations to discrete time volatility models.

1 Stochastic Volatility Models and Discrete GARCH

Both stochastic volatility models and GARCH processes are popular models for the description of financial time series. Recall that a *discrete time stochastic volatility model* (SV-*model*) is a process $(X_n)_{n\in\mathbb{N}_0}$ together with a non-negative *volatility process* $(\sigma_n)_{n\in\mathbb{N}_0}$, such that

$$X_n = \sigma_n \varepsilon_n, \quad n \in \mathbb{N}_0, \tag{1}$$

where the *noise sequence* $(\varepsilon_n)_{n\in\mathbb{N}_0}$ is a sequence of independent and identically distributed (i.i.d.) random variables, which is assumed to be *independent of* $(\sigma_n)_{n\in\mathbb{N}_0}$. Further information about these processes can be found e.g. in Shephard and Andersen (2008) and Davis and Mikosch (2008). In contrast to stochastic volatility models, GARCH processes have the property that the volatility process is specified as a function of the past observations. The classical ARCH(1) process by Engle (1982) and the GARCH(1,1) process by Bollerslev (1986), for example, are processes $(X_n)_{n\in\mathbb{N}_0}$ with a non-negative *volatility process* $(\sigma_n)_{n\in\mathbb{N}_0}$, such that

Alexander M. Lindner
Technische Universität Braunschweig, Institut für Mathematische Stochastik, Pockelsstraße 14, D-38106 Braunschweig, Germany, e-mail: a.lindner@tu-bs.de

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_21,

$$X_n = \sigma_n \varepsilon_n, \quad n \in \mathbb{N}_0, \tag{2}$$

$$\sigma_n^2 = \omega + \lambda X_{n-1}^2 + \delta \sigma_{n-1}^2, \quad n \in \mathbb{N}. \tag{3}$$

Here, $(\varepsilon_n)_{n\in\mathbb{N}_0}$ is again an i.i.d. noise sequence, and the parameters ω, λ, δ satisfy $\omega > 0$, $\lambda > 0$ and $\delta > 0$ (GARCH(1,1)) or $\delta = 0$ (ARCH(1)), respectively. See e.g. Teräsvirta (2008) and Lindner (2008) for further information regarding GARCH processes and their probabilistic properties.

While financial data are usually observed only at discrete times, financial mathematicians often tend to work in continuous time, which appears to be more convenient for option pricing. However, continuous time models may also offer a good approximation to discrete observations. Typical examples are high-frequency data or irregularly observed data. While in discrete time, $(X_n)_{n\in\mathbb{N}_0}$ models the increments of the log price, in continuous time one rather models the log price $(G_t)_{t\geq 0}$ itself. Typically, one has an unobserved volatility process $(\sigma_t)_{t\geq 0}$ modelled as a semimartingale, and the log price is described by

$$G_t = \int_0^t (\mu + b\sigma_s^2)\, ds + \int_0^t \sigma_{s-} dM_s, \tag{4}$$

where $(M_t)_{t\geq 0}$ is a Lévy process and μ, b are real constants. For *continuous time stochastic volatility models*, the process $(M_t)_{t\geq 0}$ is usually independent of $(\sigma_t)_{t\geq 0}$, and more specifically, taken to be a standard Brownian motion. In the latter case, the quadratic variation of G until time t is $\int_0^t \sigma_s^2\, ds$, justifying the name *volatility* for σ_t.

The aim of this paper is to present some continuous time GARCH and SV models and to discuss in which sense they can be seen as approximations to corresponding discrete time models.

2 Continuous Time GARCH Approximations

If a continuous time model serves as an approximation to a GARCH process, one may ask in which sense the process when sampled at discrete times is close to a GARCH process. An optimal situation would be that the process itself is a GARCH process, whenever sampled at equidistant times $(hn)_{n\in\mathbb{N}_0}$, for each $h > 0$. This however cannot be achieved: Drost and Nijman (1993), Example 3, have shown that GARCH processes are not closed under temporal aggregation, i.e., if $(X_t)_{t\in\mathbb{N}_0}$ is a GARCH process driven by some noise $(\varepsilon_t)_{t\in\mathbb{N}_0}$ with volatility process $(\sigma_t)_{t\in\mathbb{N}_0}$, then – apart from some situations when the noise is degenerate – there does not exist an i.i.d. noise sequence $(\widetilde{\varepsilon}_{2t})_{t\in\mathbb{N}_0}$ and a volatility process $(\widetilde{\sigma}_{2t})_{t\in\mathbb{N}_0}$ such that $(X_{2t})_{t\in\mathbb{N}_0}$ is a GARCH process driven by $(\widetilde{\varepsilon}_{2t})_{t\in\mathbb{N}_0}$ with volatility process $(\widetilde{\sigma}_{2t})_{t\in\mathbb{N}_0}$. In particular, a continuous time process $(Y_t)_{t\geq 0}$ which happens to be a GARCH(1,1) process when sampled at $\{0, h, 2h, \ldots\}$ for some frequency h will not be GARCH when

sampled at $\{0, 2h, 4h, \ldots\}$. Similarly, if for some log price process $(G_t)_{t\geq 0}$, the increments $(G_{nh} - G_{(n-1)h})_{n\in\mathbb{N}}$ of length h constitute a GARCH(1,1) process with non-degenerate noise, then the increments $(G_{n2h} - G_{(n-1)2h})_{n\in\mathbb{N}}$ of length $2h$ will usually not be GARCH. Hence one has to work with other concepts of GARCH approximations.

2.1 Preserving the random recurrence equation property

One approach to construct continuous time GARCH approximations is to require that certain properties of discrete time GARCH continue to hold. For example, a GARCH(1,1) process has the elegant property that its squared volatility satisfies the random recurrence equation $\sigma_n^2 = \omega + (\delta + \lambda\varepsilon_{n-1}^2)\sigma_{n-1}^2$, where σ_{n-1} is independent of ε_{n-1}. Denoting $A_n^{n-1} = \delta + \lambda\varepsilon_{n-1}^2$ and $C_n^{n-1} = \omega$, this can be written as

$$\sigma_n^2 = A_n^{n-1}\sigma_{n-1}^2 + C_n^{n-1},$$

where σ_{n-1}^2 is independent of (A_n^{n-1}, C_n^{n-1}) and $(A_n^{n-1}, C_n^{n-1})_{n\in\mathbb{N}}$ is i.i.d. Requiring at least this random recurrence equation property to hold for candidates of squared volatility processes, it is natural to look at processes $(Y_t)_{t\geq 0}$ which satisfy

$$Y_t = A_t^s Y_s + C_t^s, \quad 0 \leq s \leq t, \tag{5}$$

for appropriate sequences $(A_t^s, C_t^s)_{0\leq s\leq t}$ of bivariate random vectors. In order to ensure the i.i.d. property of $(A_{nh}^{(n-1)h}, C_{nh}^{(n-1)h})_{n\in\mathbb{N}}$ for every $h > 0$, one rather assumes that for every $0 \leq a \leq b \leq c \leq d$, the families of random variables $(A_t^s, C_t^s)_{a\leq s\leq t\leq b}$ and $(A_t^s, C_t^s)_{c\leq s\leq t\leq d}$ are independent, and that the distribution of $(A_{t+h}^{s+h}, C_{t+h}^{s+h})_{0\leq s\leq t}$ does not depend on $h \geq 0$. Finally, a natural continuity condition seems to be desirable, namely that

$$A_t^0 > 0 \text{ a.s.} \quad \forall\, t \geq 0, \quad \text{and} \quad (A_t, C_t) := (A_t^0, C_t^0) \xrightarrow{P} (1, 0) \quad \text{as } t \downarrow 0,$$

where "$\xrightarrow{P}$" denotes convergence in probability. De Haan and Karandikar (1989) showed that if $(A_t^s, C_t^s)_{0\leq s\leq t}$ are such that they satisfy the properties described above, then $(A_t, C_t)_{t\geq 0}$ admit càdlàg versions, and with these versions chosen, $(Y_t)_{t\geq 0}$ satisfies (5) if and only if there is a bivariate Lévy process $(\xi_t, \eta_t)_{t\geq 0}$ such that

$$Y_t = e^{-\xi_t}\left(Y_0 + \int_0^t e^{\xi_{s-}}\, d\eta_s\right), \quad t \geq 0. \tag{6}$$

This is a *generalised Ornstein-Uhlenbeck process*, which is discussed in detail in Maller et al. (2008b). From the point of view described above, generalised Ornstein-Uhlenbeck processes are natural continuous time analogues

of random recurrence equations, and hence a desirable property of a continuous time GARCH(1,1) approximation is that its squared volatility process is a generalised Ornstein-Uhlenbeck process. As we shall see later, both the diffusion limit of Nelson (1990) as well as the COGARCH(1,1) process of Klüppelberg et al. (2004) satisfy this requirement. Also, the volatility model of Barndorff-Nielsen and Shephard (2001a) and (2001b) falls into this class, even if not constructed as a GARCH(1,1) approximation.

2.2 The diffusion limit of Nelson

A common method to construct continuous time processes from discrete ones is to use a diffusion approximation. Here, one takes a series of discrete time series defined on a grid (such as $h\mathbb{N}_0$) with mesh $h \downarrow 0$, extends the processes between grid points in a suitable way (such as interpolation, or piecewise constancy), and hopes that this sequence of processes defined on $[0,\infty)$ converges weakly to some limit process. Since the processes encountered will typically have sample paths in the Skorokhod space $D([0,\infty),\mathbb{R}^d)$ of $\mathbb{R}^d$-valued càdlàg functions defined on $[0,\infty)$, by *weak convergence* we mean weak convergence in $D([0,\infty),\mathbb{R}^d)$, when endowed with the (J_1-)Skorokhod topology, cf. Jacod and Shiryaev (2003), Sections VI.1 and VI.3. If the limit process has no fixed points of discontinuity (which will be the case in all cases encountered), then weak convergence in $D([0,\infty),\mathbb{R}^d)$ implies weak convergence of the finite dimensional distributions, and the converse is true under an additional tightness condition, cf. Jacod and Shiryaev (2003), Proposition VI.3.14 and VI.3.20.

Nelson (1990) derived a diffusion limit for GARCH(1,1) processes. In the same paper, he also considered the diffusion limit of EGARCH processes. An extension to diffusion limits of a more general class of GARCH processes (called *augmented* GARCH) was obtained by Duan (1997). Here, we shall concentrate on Nelson's diffusion limit of GARCH(1,1): for each $h > 0$, let $(\varepsilon_{kh,h})_{k\in\mathbb{N}_0}$ be an i.i.d. sequence of standard normal random variables, let $\omega_h, \lambda_h > 0$ and $\delta_h \geq 0$, and let $(G_{0,h}, \sigma^2_{0,h})$ be starting random variables, independent of $(\varepsilon_{kh,h})_{k\in\mathbb{N}_0}$. Then $(G_{kh,h} - G_{(k-1)h,h}, \sigma_{kh,h})_{k\in\mathbb{N}}$, defined recursively by

$$\begin{aligned} G_{kh,h} &= G_{(k-1)h,h} + h^{1/2}\sigma_{kh,h}\varepsilon_{kh,h}, \quad k \in \mathbb{N}, \\ \sigma^2_{kh,h} &= \omega_h + (\lambda_h \varepsilon^2_{(k-1)h,h} + \delta_h)\sigma^2_{(k-1)h,h}, \quad k \in \mathbb{N}, \end{aligned}$$

is a GARCH(1,1) process for every $h > 0$. Then $(G_{kh,h}, \sigma^2_{kh,h})_{k\in\mathbb{N}_0}$ is embedded into a continuous time process $(G_{t,h}, \sigma^2_{t,h})_{t\geq 0}$ by defining

$$G_{t,h} := G_{kh,h}, \quad \sigma^2_{t,h} := \sigma^2_{kh,h}, \quad kh \leq t < (k+1)h.$$

The latter process has sample paths in $D([0,\infty),\mathbb{R}^2)$, and Nelson (1990) gives conditions for $(G_{t,h},\sigma^2_{t,h})_{t\geq 0}$ to converge weakly to some process $(G_t,\sigma^2_t)_{t\geq 0}$ as $h\downarrow 0$. Namely, suppose that there are constants $\omega\geq 0$, $\theta\in\mathbb{R}$ and $\lambda>0$ as well as starting random variables (G_0,σ^2_0) such that $(G_{0,h},\sigma^2_{0,h})$ converges weakly to (G_0,σ^2_0) as $h\downarrow 0$, such that $P(\sigma^2_0>0)=1$ and

$$\lim_{h\downarrow 0} h^{-1}\omega_h=\omega,\quad \lim_{h\downarrow 0} h^{-1}(1-\delta_h-\lambda_h)=\theta,\quad \lim_{h\downarrow 0} 2h^{-1}\lambda_h^2=\lambda^2. \tag{7}$$

Then $(G_{t,h},\sigma^2_{t,h})_{t\geq 0}$ converges weakly as $h\downarrow 0$ to the unique solution $(G_t,\sigma^2_t)_{t\geq 0}$ of the diffusion equation

$$dG_t=\sigma_t\,dB_t,\quad t\geq 0, \tag{8}$$

$$d\sigma^2_t=(\omega-\theta\sigma^2_t)\,dt+\lambda\sigma^2_t dW_t,\quad t\geq 0, \tag{9}$$

with starting value (G_0,σ^2_0), where $(B_t)_{t\geq 0}$ and $(W_t)_{t\geq 0}$ are independent Brownian motions, independent of (G_0,σ^2_0). Nelson also showed that (9) has a strictly stationary solution $(\sigma^2_t)_{t\geq 0}$ if $2\theta/\lambda^2>-1$ and $\omega>0$, in which case the marginal stationary distribution of σ^2_0 is inverse Gamma distributed with parameters $1+2\theta/\lambda^2$ and $2\omega/\lambda^2$. An example for possible parameter choices to satisfy (7) is given by $\omega_h=\omega h$, $\delta_h=1-\lambda\sqrt{h/2}-\theta h$, and $\lambda_h=\lambda\sqrt{h/2}$.

Observe that the limit volatility process $(\sigma^2_t)_{t\geq 0}$ in (9) is a generalised Ornstein-Uhlenbeck process as defined in (6), with $(\xi_t,\eta_t)=(-\lambda W_t+(\theta+\lambda^2/2)t,\omega t)$, see e.g. Fasen (2008) or Maller et al. (2008b).

A striking difference between discrete time GARCH processes and their diffusion limit is that the squared volatility process $(\sigma^2_t)_{t\geq 0}$ in (9) is independent of the Brownian motion $(B_t)_{t\geq 0}$, driving the log price process. So the volatility model (8), (9) has two independent sources of randomness, namely $(B_t)_{t\geq 0}$ and $(W_t)_{t\geq 0}$. On the other hand, discrete time GARCH processes are defined only in terms of a single noise sequence $(\varepsilon_n)_{n\in\mathbb{N}_0}$, rather than two. While it was believed for a long time that Nelson's limit result justified the estimation of stochastic volatility models by GARCH-estimation procedures, Wang (2002) showed that statistical inference for GARCH modelling and statistical inference for the diffusion limit (8), (9) are not asymptotically equivalent, where asymptotic equivalence is defined in terms of Le Cam's deficiency distance (see Le Cam (1986)). As a heuristic explanation, Wang (2002) mentions the different kinds of noise propagations in the GARCH model with one source of randomness and the volatility model with two sources of randomness. It is possible to modify Nelson's approximation to obtain a limit process which is driven by a single Brownian motion only (see Corradi (2000)), but in that case the limiting volatility process is deterministic, an undesirable property of price processes. Observe however that for the latter case, the statistical estimation procedures are equivalent, cf. Wang (2002).

2.3 The COGARCH model

Apart from the fact that Nelson's diffusion limit is driven by two independent sources of randomness, it also has a continuous volatility process. Nowadays, jumps in the volatility of continuous time processes are often considered as a stylised fact, which hence is not met by model (8), (9). This led Klüppelberg et al. (2004) to the introduction of a new continuous time GARCH(1,1) process, called COGARCH(1,1), where "CO" stands for continuous time. Its construction starts from the observation that the recursions (2), (3) can be solved recursively, and σ_n^2 and X_n can be expressed by

$$\sigma_n^2 = \omega \sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} (\delta + \lambda\varepsilon_j^2) + \sigma_0^2 \prod_{j=0}^{n-1} (\delta + \lambda\varepsilon_j^2) \tag{10}$$

$$= \left(\omega \int_0^n \exp\left\{ -\sum_{j=0}^{\lfloor s \rfloor} \log(\delta + \lambda\varepsilon_j^2) \right\} ds + \sigma_0^2 \right) \exp\left\{ \sum_{j=0}^{n-1} \log(\delta + \lambda\varepsilon_j^2) \right\},$$

$$X_n = \sigma_n \varepsilon_n = \sigma_n \left(\sum_{j=0}^{n} \varepsilon_j - \sum_{j=0}^{n-1} \varepsilon_j \right). \tag{11}$$

Here, $\lfloor z \rfloor$ denotes the largest integer not exceeding z, and both $\sum_{j=0}^{n-1} \log(\delta + \lambda\varepsilon_j^2) = n \log \delta + \sum_{j=0}^{n-1} \log(1 + \lambda\varepsilon_j^2/\delta)$ and $\sum_{j=0}^{n} \varepsilon_j$ are random walks, which are linked in such a way that the first can be reconstructed from the second. The idea of Klüppelberg et al. (2004) was then to replace the appearing random walks by Lévy processes, which are a continuous time analogue of random walks, and to replace the ε_j by the jumps of a Lévy process L. More precisely, they start with positive constants $\omega, \delta, \lambda > 0$ and a Lévy process $L = (L_t)_{t \geq 0}$ which has non-zero Lévy measure ν_L, and define an auxiliary Lévy process ξ_t by

$$\xi_t = -t \log \delta - \sum_{0 < s \leq t} \log\left(1 + \frac{\lambda}{\delta} (\Delta L_s)^2 \right), \quad t \geq 0,$$

corresponding to the random walk $-(n \log \delta + \sum_{j=0}^{n-1} \log(1 + \lambda\varepsilon_j^2/\delta))$ in discrete time. Given a starting random variable σ_0^2, independent of $(L_t)_{t \geq 0}$, a *(right-continuous) volatility process* $(\sigma_t)_{t \geq 0}$ and the COGARCH(1,1) process are then defined by

$$\sigma_t^2 = \left(\omega \int_0^t e^{\xi_{s-}} \, ds + \sigma_0^2 \right) e^{-\xi_t}, \quad t \geq 0, \tag{12}$$

$$G_t = \int_0^t \sigma_{s-} \, dL_s, \quad t \geq 0, \tag{13}$$

in complete analogy to (10) and (11). (Originally, Klüppelberg et al. (2004) defined a *left-continuous* version of the volatility by considering σ_{t-}^2 rather than σ_t^2.) The process $(\xi_t)_{t\geq 0}$ is indeed a Lévy process, which is the negative of a subordinator together with drift $-\log\delta$. Hence the squared volatility process is again a generalised Ornstein-Uhlenbeck process as in (6) driven by $(\xi_t, \omega t)$. An application of Itô's formula to (12) shows that $(\sigma_t^2)_{t\geq 0}$ satisfies the stochastic differential equation

$$d\sigma_t^2 = (\omega + \log\delta\,\sigma_{t-}^2)\,dt + \frac{\lambda}{\delta}\,\sigma_{t-}^2\,d[L,L]_t^d,$$

where $[L,L]_t^d = \sum_{0<s\leq t}(\Delta L_s)^2$ denotes the discrete part of the quadratic variation of L. Note that G has only one source of randomness, namely L, which drives both σ_t^2 and G. In particular, if L jumps then so does G, with jump size $\Delta G_t = \sigma_{t-}\Delta L_t$. Stationarity and moment conditions for $(\sigma_t^2)_{t\geq 0}$ are given in Klüppelberg et al. (2004), and it follows that $(\sigma_t^2)_{t\geq 0}$ admits a stationary version if and only if $\int_{\mathbb{R}} \log(1+\lambda x^2/\delta)\,\nu_L(dx) < -\log\delta$, which in particular forces $\delta < 1$. As for discrete time GARCH processes, the stationary COGARCH volatility has Pareto tails under weak assumptions, cf. Klüppelberg et al. (2006). Under appropriate conditions, the increments of G are uncorrelated, while the squares of the increments are correlated. More precisely, the covariance structure of $((G_{nh} - G_{(n-1)h})^2)_{n\in\mathbb{N}}$ is that of an ARMA(1,1) process. Extensions of the COGARCH(1,1) process include the COGARCH(p,q) model by Brockwell et al. (2006), an asymmetric COGARCH(1,1) model by Haug et al. (2007) to include the leverage effect, and a multivariate COGARCH(1,1) model by Stelzer (2008).

The COGARCH(1,1) model was motivated by replacing the innovations in GARCH processes by the jumps of Lévy processes. The question in which sense a COGARCH process is close to a discrete time GARCH process was recently considered independently by Kallsen and Vesenmayer (2008) as well as Maller et al. (2008a). In both papers it is shown that the COGARCH(1,1) model is a continuous time limit of certain GARCH(1,1) processes: more precisely, given a COGARCH process $(G_t)_{t\geq 0}$ with volatility process $(\sigma_t)_{t\geq 0}$, Kallsen and Vesenmayer (2008) construct a sequence of discrete time GARCH processes $(Y_{k,n})_{k\in\mathbb{N}}$ with volatility $(\sigma_{k,n})_{k\in\mathbb{N}}$, such that the processes $(\sum_{k=1}^{\lfloor nt\rfloor} Y_{k,n}, \sigma_{\lfloor nt\rfloor+1,n})_{t\geq 0}$ converge weakly to $(G_t,\sigma_t)_{t\geq 0}$ as $n\to\infty$. Here, weak convergence in the Skorokhod space $D([0,\infty),\mathbb{R}^2)$ is obtained by computing the semimartingale characteristics of $(G_t,\sigma_t)_{t\geq 0}$ and showing that they are the limit of those of $(\sum_{k=1}^{\lfloor nt\rfloor} Y_{k,n}, \sigma_{\lfloor nt\rfloor+1,n})_{t\geq 0}$ as $n\to\infty$. The infinitesimal generator of the strong Markov process $(G_t,\sigma_t)_{t\geq 0}$ was also obtained. They also showed how a given GARCH(1,1) process can be scaled to converge to a COGARCH(1,1) process. Using completely different methods, given a COGARCH(1,1) process driven by a Lévy process with mean zero and finite variance, Maller et al. (2008a) also obtain a sequence of discrete time GARCH processes $(Y_{k,n})_{k\in\mathbb{N}}$ with volatility processes $(\sigma_{k,n})_{k\in\mathbb{N}}$

such that $(\sum_{k=1}^{\lfloor nt \rfloor} Y_{k,n}, \sigma_{\lfloor nt \rfloor+1,n})_{t\geq 0}$ converges in probability to $(G_t, \sigma_t)_{t\geq 0}$ as $n \to \infty$. Observe that convergence in probability is stronger than weak convergence. The discrete time GARCH processes are constructed using a "first-jump" approximation for Lévy processes as developed by Szimayer and Maller (2007), which divides a compact interval into an increasing number of subintervals and for each subinterval takes the first jump exceeding a certain threshold. Summing up, we have seen that the COGARCH(1,1) model is a limit of GARCH(1,1) processes, although originally motivated by mimicking features of discrete GARCH(1,1) processes without referring to limit procedures.

2.4 Weak GARCH processes

Another approach to obtain continuous time GARCH processes is to weaken the definition of a GARCH process. Observe that if $(X_n)_{n\in\mathbb{N}_0}$ is a GARCH(1,1) process with finite fourth moment and volatility process $(\sigma_n)_{n\in\mathbb{N}_0}$, driven by i.i.d. noise $(\varepsilon_n)_{n\in\mathbb{N}_0}$ such that $E\varepsilon_0 = 0$ and $E\varepsilon_0^2 = 1$, then

$$PL_n(X_n) = 0 \quad \text{and} \quad PL_n(X_n^2) = \sigma_n^2,$$

where $PL_n(Z)$ denotes the best linear predictor of a square integrable random variable Z with respect to 1, σ_0^2, X_0, …, X_{n-1}, X_0^2, …, X_{n-1}^2. Drost and Nijman (1993) use this property to define weak GARCH processes: they call a univariate process $(X_n)_{n\in\mathbb{N}_0}$ a *weak* GARCH(1,1) process with parameter $(\omega, \lambda, \delta)$, if X_n has finite fourth moment and there exists a volatility process $(\sigma_n)_{n\in\mathbb{N}_0}$ such that $(\sigma_n^2)_{n\in\mathbb{N}_0}$ is weakly stationary and satisfies (3) for $n \in \mathbb{N}$, and it holds $PL_n(X_n) = 0$ and $PL_n(X_n^2) = \sigma_n^2$. Here, $\omega > 0$, $\lambda \geq 0$, $\delta \geq 0$, and either $\lambda = \delta = 0$ or $0 < \lambda + \delta < 1$. Unlike GARCH processes, the class of weak GARCH processes is closed under temporal aggregation, i.e. if $(X_n)_{n\in\mathbb{N}_0}$ is a symmetric weak GARCH(1,1) process, then so is $(X_{mn})_{n\in\mathbb{N}_0}$ for every $m \in \mathbb{N}$, see Drost and Nijman (1993), Example 1. Based on this property, Drost and Werker (1996) define a *continuous time weak* GARCH(1,1) *process* to be a univariate process $(G_t)_{t\geq 0}$ such that $(G_{t_0+nh} - G_{t_0+(n-1)h})_{n\in\mathbb{N}}$ is a weak GARCH(1,1) process for every $h > 0$ and $t_0 \geq 0$. They also show that the parameters of the discretised weak GARCH process correspond to certain parameters in the continuous time weak GARCH process, so that estimation methods for discrete time weak GARCH processes carry over to certain parameters of continuous time weak GARCH processes. Examples of continuous time weak GARCH processes include the diffusion limit of Nelson, provided it has finite fourth moment, or more generally processes $(G_t)_{t\geq 0}$ with finite fourth moments of the form $dG_t = \sigma_{t-} dL_t$, where $(\sigma_t^2)_{t\geq 0}$ is supposed to be a stationary solution of the stochastic differential equation

$$d\sigma_t^2 = (\omega - \theta\sigma_{t-}^2)\,dt + \lambda\sigma_{t-}^2\,d\eta_t.$$

Here, $(L_t)_{t\geq 0}$ and $(\eta_t)_{t\geq 0}$ are two independent Lévy processes with finite fourth moment, expectation 0 and variance 1, $(L_t)_{t\geq 0}$ is symmetric and the parameters $\omega > 0$, $\theta > 0$ and $\lambda < 1$ are chosen such that $E\sigma_0^4 < \infty$, see Drost and Werker (1996), Example 4.1.

2.5 Stochastic delay equations

A somewhat different approach to obtain continuous time GARCH processes is taken by Lorenz (2006). He considered a weak limit of scaled GARCH$(pn+1,1)$ processes when the order $pn+1$ goes to ∞, and in the limit he obtained the solution to a stochastic delay differential equation. More precisely, let $(\varepsilon_k)_{k\in\mathbb{N}_0}$ be a sequence of i.i.d. random variables with finite $(4+\alpha)$-moment for some $\alpha > 0$ such that $E(\varepsilon_1) = E(\varepsilon_1^3) = 0$ and $E(\varepsilon_1^2) = 1$. Let $p \in \mathbb{N}$ and $(\sigma_t)_{t\in[-p,0]}$ be some given strictly positive continuous function on $[-p,0]$, and define $G_t = 0$ for $t \in [-p,0]$. Let $\omega_n > 0$, $\delta_{j,n} \geq 0$ $(j = 0,\ldots,pn)$ and $\lambda_n \geq 0$, and consider the discrete time GARCH$(pn+1,1)$ process $(Y_{k,n})_{k\in\mathbb{N}}$ with volatility $(\sigma_{k,n})_{k\in\mathbb{N}}$ given by

$$\begin{aligned} Y_{k,n} &= n^{-1/2}\,\sigma_{k,n}\varepsilon_k, \quad k \in \mathbb{N}, \\ \sigma_{k,n}^2 &= \omega_n + \sum_{j=0}^{np} \delta_{j,n}\sigma_{k-1-j,n}^2 + \lambda_n\sigma_{k-1,n}^2\varepsilon_{k-1}^2, \quad k \in \mathbb{N}, \end{aligned}$$

where $\sigma_{j,n} := \sigma_{-j/n}$ for $j \in \{-pn,\ldots,0\}$. Define further $(G_{t,n},\sigma_{t,n})_{t\geq -p} := (\sum_{k=1}^{\lfloor nt\rfloor} Y_{k,n}, \sigma_{\lfloor nt\rfloor+1,n})_{t\geq -p}$. Assuming that

$$\lim_{n\to\infty} n\omega_n = \omega > 0,\ \lim_{n\to\infty} n(1-\delta_{0,n}-\lambda_n) = \theta \in \mathbb{R},\ \lim_{n\to\infty}(E\varepsilon_1^4 - 1)n\lambda_n^2 = \lambda^2 \tag{14}$$

and that the sequence $(\gamma_n)_{n\in\mathbb{N}}$ of discrete measures γ_n on $[-p,0]$ defined by $\gamma_n(\{-j/n\}) = n\delta_{j,n}$ for $1 \leq j \leq pn$ and $\gamma_n(\{0\}) = 0$ converges vaguely to some finite measure γ on $[-p,0]$ such that $\gamma(\{0\}) = 0$, Lorenz (2006), Theorem 2.5.10, showed that $(G_{t,n},\sigma_{t,n})_{t\geq 0}$ converges weakly as $n \to \infty$ to the unique weak solution $(G_t,\sigma_t)_{t\geq 0}$ of the stochastic delay differential equation

$$dG_t = \sigma_t\,dB_t, \quad t \geq 0, \tag{15}$$

$$d\sigma_t^2 = (\omega - \theta\sigma_t^2)\,dt + \left(\int_{[-p,0]} \sigma_{t+u}^2\,d\gamma(u)\right)dt + \lambda\sigma_t^2\,dW_t, \tag{16}$$

with starting values as given, and where $(B_t)_{t\geq 0}$ and $(W_t)_{t\geq 0}$ are two independent Brownian motions. A sufficient condition for a stationary solution

of the stochastic delay equation (16) to exist is also given in Lorenz (2006). Observe that if $\delta_{j,n} = 0$ for $j = 1, \ldots, pn$, the discrete GARCH$(pn+1,1)$ processes are actually GARCH(1,1) processes, the limit measure γ is zero, and (14), (15), (16) reduce to the corresponding equations (7), (8), (9) for Nelson's diffusion limit.

A related paper regarding limits of HARCH processes which give rise to stochastic delay equations is Zheng (2005).

2.6 A continuous time GARCH model designed for option pricing

The previous continuous time GARCH models have been mainly designed as limits of discrete time GARCH processes or as processes with properties similar to GARCH. Option pricing for such models may be demanding, since they often give rise to incomplete markets. Inspired by this, Kallsen and Taqqu (1998) developed a continuous time process which is a GARCH process when sampled at integer times. Their process is also driven by a single Brownian motion only. More specifically, let $\omega, \lambda > 0$, $\delta \geq 0$ and $(B_t)_{t\geq 0}$ be a standard Brownian motion. For some starting random variable σ_0^2, define the volatility process $(\sigma_t)_{t\geq 0}$ by $\sigma_t^2 = \sigma_0^2$ for $t \in [0,1)$ and

$$\sigma_t^2 = \omega + \lambda \left(\int_{\lfloor t \rfloor - 1}^{\lfloor t \rfloor} \sigma_{s-} dB_s \right)^2 + \delta \sigma_{\lfloor t \rfloor - 1}^2, \quad t \geq 1. \tag{17}$$

The continuous time GARCH process $(G_t)_{t\geq 0}$ then models the log-price process, and is given by

$$G_t = G_0 + \int_0^t (\mu(\sigma_{s-}) - \sigma_{s-}^2/2)\, ds + \int_0^t \sigma_s \, dB_s.$$

Here, the drift function μ is assumed to have continuous derivatives. Observe that the volatility process $(\sigma_t)_{t\geq 0}$ given by (17) is constant on intervals $[n, n+1)$ for $n \in \mathbb{N}_0$. Also observe that the process $(G_t - G_{t-1}, \sigma_{t-1})_{t\geq 1}$, when sampled at integer times, gives rise to a discrete time GARCH(1,1)-M process

$$\begin{aligned} G_n - G_{n-1} &= \mu(\sigma_{n-1}) - \sigma_{n-1}^2/2 + \sigma_{n-1}(B_n - B_{n-1}), \quad n \in \mathbb{N}, \\ \sigma_n^2 &= \omega + \lambda \sigma_{n-1}^2 (B_n - B_{n-1})^2 + \delta \sigma_{n-1}^2, \quad n \in \mathbb{N}. \end{aligned}$$

This differs from a usual GARCH(1,1) process only by the term $\mu(\sigma_{n-1}) - \sigma_{n-1}^2/2$, which vanishes if the function μ is chosen as $\mu(x) = x^2/2$. If we are not in the classical GARCH situation but rather have $\limsup_{x\to\infty} \mu(x)/x < \infty$, then Kallsen and Taqqu (1998) show that the continuous time model is

arbitrage free and complete. This is then used to derive pricing formulas for contingent claims such as European options.

3 Continuous Time Stochastic Volatility Approximations

Recall from Section 1 that by a discrete time stochastic volatility model we mean a process $(X_n)_{n\in\mathbb{N}_0}$ satisfying (1), where $(\varepsilon_n)_{n\in\mathbb{N}_0}$ is i.i.d. and $(\sigma_n)_{n\in\mathbb{N}_0}$ is a stochastic volatility process, independent of $(\varepsilon_n)_{n\in\mathbb{N}_0}$. Here, we shall usually restrict ourselves to the case when $(\varepsilon_n)_{n\in\mathbb{N}_0}$ is i.i.d. normally distributed with expectation zero. Also recall that we defined continuous time stochastic volatility models by (4). Now, we shall further restrict ourselves to the case where $\mu = b = 0$ and M in (4) is Brownian motion, i.e. we consider models of the form

$$G_t = \int_0^t \sigma_{s-}\, dB_s, \quad t \geq 0, \tag{18}$$

where $B = (B_t)_{t\geq 0}$ is a standard Brownian motion, independent of the volatility process $\sigma = (\sigma_t)_{t\geq 0}$. The latter is assumed to be a strictly positive semimartingale, in particular it has càdlàg paths.

3.1 Sampling a continuous time SV model at equidistant times

In the setting as given above, it is easy to see that discrete time SV models are closed under temporal aggregation, however, with a possibly unfamiliar volatility process after aggregation. Similarly, the continuous time SV model (18) gives rise to a discrete time SV model, when sampled at equidistant time points. To see the latter, let G be given by (18), $h > 0$, and define

$$\varepsilon_k := \frac{G_{kh} - G_{(k-1)h}}{(\int_{(k-1)h}^{kh} \sigma_s^2\, ds)^{1/2}}, \quad k \in \mathbb{N}.$$

Since conditionally on $(\sigma_t)_{t\geq 0}$, $G_{kh} - G_{(k-1)h}$ is normally distributed with expectation zero and variance $\int_{(k-1)h}^{kh} \sigma_s^2\, ds$, it follows that conditionally on $(\sigma_t)_{t\geq 0}$, ε_k is standard normally distributed, and since this distribution does not depend on $(\sigma_t)_{t\geq 0}$, ε_k itself is $N(0,1)$ distributed. With similar arguments one sees that ε_k is independent of σ and that $(\varepsilon_k)_{k\in\mathbb{N}}$ is i.i.d. Then $(G_{kh} - G_{(k-1)h} = \widetilde{\sigma}_k \varepsilon_k)_{k\in\mathbb{N}}$ is a discrete time stochastic volatility model, with discrete time volatility process

$$\widetilde{\sigma}_k := \left(\int_{(k-1)h}^{kh} \sigma_s^2 \, ds \right)^{1/2}, \quad k \in \mathbb{N}. \tag{19}$$

Hence we see that unlike GARCH processes, continuous time SV models yield discrete time SV models when sampled, i.e. they stay in their own class. Unfortunately, the volatility process $(\widetilde{\sigma}_k)_{k\in\mathbb{N}}$ obtained by this method is not always in a very tractable form, and often it might be desirable to retain a particular structure on the volatility process. As an illustration of a process, where most but not all of the structure is preserved, consider the stochastic volatility model of Barndorff-Nielsen and Shephard (2001a) and (2001b). Here, the volatility process $(\sigma_t)_{t\geq 0}$ is modeled via $d\sigma_t^2 = -\lambda\sigma_t^2 \, dt + dL_{\lambda t}$, where $\lambda > 0$ and L is a subordinator, i.e. a Lévy process with increasing sample paths. The solution to this Lévy driven Ornstein-Uhlenbeck process is given by

$$\sigma_t^2 = (\sigma_0^2 + \int_0^t e^{\lambda s} \, dL_{\lambda s}) e^{-\lambda t}, \quad t \geq 0. \tag{20}$$

Taking $\lambda = 1$ and $h = 1$ for simplicity, it follows that σ_t^2 satisfies

$$\sigma_t^2 = e^{-1}\sigma_{t-1}^2 + \int_{t-1}^t e^{u-t} \, dL_u, \quad t \geq 1, \tag{21}$$

so that

$$\widetilde{\sigma}_k^2 = \int_{k-1}^k \sigma_s^2 \, ds = e^{-1}\widetilde{\sigma}_{k-1}^2 + \int_{k-1}^k \int_{s-1}^s e^{u-s} \, dL_u \, ds, \quad k \in \mathbb{N} \setminus \{1\}. \tag{22}$$

Like the Ornstein-Uhlenbeck process, which is a continuous time AR(1) process, (22) yields a discrete time AR(1) process for $(\widetilde{\sigma}_k^2)_{k\in\mathbb{N}}$. However, (20) is driven by a Lévy process which can be interpreted as a continuous time analogue to i.i.d. noise, while (22) has 1-dependent noise given by $(\int_{k-1}^k \int_{s-1}^s e^{u-s} \, dL_u \, ds)_{k\in\mathbb{N}\setminus\{1\}}$.

We have seen that sampled continuous time SV models give rise to discrete time SV models, however the discrete time volatility may lose certain structural features. Allowing more general definitions of discrete and continuous time stochastic volatility models, in a spirit similar to the weak GARCH processes by Drost and Nijman (1993) and Drost and Werker (1996), Meddahi and Renault (2004) consider many continuous time SV models which keep the same structure when sampled at equidistant time points. We do not go into further detail, but refer to Meddahi and Renault (2004) and the overview article by Ghysels et al. (1996).

3.2 Approximating a continuous time SV model

Rather than working with the unfamiliar discrete time volatility $(\widetilde{\sigma}_k)_{k\in\mathbb{N}}$ as given in (19), one might try to use an Euler type approximation for the process G, and sample the (unobserved) volatility process $(\sigma_t^2)_{t\geq 0}$ directly at equidistant times. More precisely, consider G_t as defined in (18), where $(\sigma_t)_{t\geq 0}$ is a strictly positive semimartingale, independent of $(B_t)_{t\geq 0}$. For $h > 0$, define

$$Y_{k,h} := \sigma_{(k-1)h}(B_{kh} - B_{(k-1)h}), \quad k \in \mathbb{N}, \tag{23}$$

where $\sigma_{(k-1)h}$ is the continuous time volatility process $(\sigma_t)_{t\geq 0}$ taken at times $(k-1)h$. Then $(Y_{k,h})_{k\in\mathbb{N}}$ defines a discrete time SV model, which approximates $(G_{kh} - G_{(k-1)h})_{k\in\mathbb{N}}$. Indeed, since $(\sigma_t(\omega))_{t\geq 0}$ is a càdlàg function for almost every ω in the underlying probability space, one can easily show that the sequence of processes $\sigma^{(n)} = \sum_{k=1}^{\infty} \sigma_{k/n} \mathbf{1}_{[(k-1)/n,k/n)}$ converges almost surely on every compact interval $[0,T]$ with $T \in \mathbb{N}$ in the Skorokhod topology of $D([0,T],\mathbb{R})$ to $(\sigma_t)_{0\leq t\leq T}$, as $n \to \infty$. On the other hand, the process $(\sum_{k=1}^{\lfloor nt\rfloor+1} Y_{k,1/n})_{0\leq t\leq T}$ converges uniformly in probability to $(\int_0^t \sigma_{s-}\, dB_s)_{0\leq t\leq T}$ as $n \to \infty$, see Protter (2004), Theorem II.21. Using the continuity of $(G_t)_{t\geq 0}$, it is then easy to deduce that the bivariate process $(\sigma^{(n)}(t), \sum_{k=1}^{\lfloor nt\rfloor+1} Y_{k,1/n})_{0\leq t\leq T}$ converges in probability to $(\sigma_t, G_t)_{0\leq t\leq T}$ in the Skorokhod space $D([0,T],\mathbb{R}^2)$, from which convergence in probability on the whole space $D([0,\infty),\mathbb{R}^2)$ can be deduced. Hence the continuous time SV model is a limit of the discrete time SV models (23), as $h = 1/n \to 0$. The structure of $(\sigma_t)_{t\geq 0}$ is usually much more compatible with the structure of $(\sigma_{kh})_{k\in\mathbb{N}}$ than with the structure of $(\widetilde{\sigma}_k)_{k\in\mathbb{N}}$ of (19), and often the discretisation (23) leads to popular discrete time SV models. We give some examples.

Example 1 In the volatility model of Hull and White (1987) the continuous time volatility process $(\sigma_t)_{t\geq 0}$ follows a geometric Brownian motion, i.e. $d\sigma_t^2 = \sigma_t^2(b\,dt + \delta\, dW_t)$, where $(W_t)_{t\geq 0}$ is a Brownian motion. Then $\sigma_t^2 = \exp\{(b - \delta^2/2)t + \delta W_t\}$, so that for each $h > 0$, $\log\sigma_{kh}^2 - \log\sigma_{(k-1)h}^2 = (b-\delta^2/2)h + \delta(W_{kh} - W_{(k-1)h})$, meaning that $(\log\sigma_{kh}^2)_{k\in\mathbb{N}_0}$ is a random walk with i.i.d. $N((b-\delta^2/2)h, \delta^2)$ innovations.

Example 2 In the volatility model of Wiggins (1987), see also Scott (1987), the log-volatility is modelled as a Gaussian Ornstein-Uhlenbeck process, i.e. σ_t^2 satisfies the stochastic differential equation $d\log\sigma_t^2 = (b_1 - b_2\log\sigma_t^2)\,dt + \delta\, dW_t$ with a Brownian motion $(W_t)_{t\geq 0}$. The solution to this equation is

$$\log\sigma_t^2 = e^{-b_2 t}\left(\log\sigma_0^2 + \int_0^t e^{b_2 s}(b_1\, ds + \delta\, dW_s)\right), \quad t \geq 0,$$

so that for each $h > 0$ we obtain

$$\log \sigma_{kh}^2 = e^{-b_2 h} \log \sigma_{(k-1)h}^2 + \int_{(k-1)h}^{kh} e^{b_2(s-kh)}(b_1\, ds + \delta\, dW_s), \quad k \in \mathbb{N},$$

which is an AR(1) process with i.i.d. normal noise. So in this case we recognize model (23) as the volatility model of Taylor (1982). Unlike for Nelson's GARCH(1,1) diffusion limit, the continuous time SV model of Wiggins and its diffusion approximation (23) are statistically equivalent, as investigated by Brown et al. (2003).

Example 3 In the volatility model of Barndorff-Nielsen and Shephard (2001a) and (2001b), where the squared volatility is modelled as a subordinator driven Ornstein-Uhlenbeck process, one obtains similarly to (21),

$$\sigma_{kh}^2 = e^{-\lambda h}\sigma_{(k-1)h}^2 + \int_{(k-1)h}^{kh} e^{\lambda(u-kh)}\, dL_{\lambda u}, \quad k \in \mathbb{N},$$

so that the discretised squared volatility satisfies an AR(1) process with non-Gaussian but positive i.i.d. noise.

Example 4 If one models the squared volatility $(\sigma_t^2)_{t\geq 0}$ by a subordinator driven continuous time ARMA process (CARMA) as suggested by Brockwell (2004), then the discretised squared volatility follows a discrete time ARMA process, but not necessarily with i.i.d. noise, see Brockwell (2008). If one models instead the log-volatility $(\log \sigma_t^2)_{t\geq 0}$ by a Lévy driven CARMA process, similarly to the method of Haug and Czado (2007) who specify the volatility of an exponential continuous time GARCH process in this way, then the discretised log-volatility $(\log \sigma_{kh}^2)_{k\in\mathbb{N}}$ follows a discrete time ARMA process. If the driving Lévy process is a Brownian motion, then the discrete time ARMA process also has i.i.d. Gaussian noise.

Example 5 If one approximates the GARCH diffusion limit (8), (9) via (23), the resulting discretised squared volatility process $(\sigma_{kh}^2)_{k\in\mathbb{N}_0}$ satisfies a random recurrence equation $\sigma_{kh}^2 = A_{kh}^{(k-1)h}\sigma_{(k-1)h}^2 + C_{kh}^{(k-1)h}$ with i.i.d. $(A_{kh}^{(k-1)h}, C_{kh}^{(k-1)h})_{k\in\mathbb{N}}$, where

$$\begin{aligned} A_{kh}^{(k-1)h} &= e^{\lambda(W_{kh}-W_{(k-1)h})-(\theta+\lambda^2/2)h}, \\ C_{kh}^{(k-1)h} &= \omega \int_{(k-1)h}^{kh} e^{\lambda(W_{kh}-W_s)-(\theta+\lambda^2/2)(kh-s)}\, ds. \end{aligned}$$

This follows from the fact that the squared volatility process satisfies a generalised Ornstein-Uhlenbeck process as pointed out in Section 2. Also observe that a random recurrence equation may be viewed as kind of an AR(1) process with random coefficients.

Summing up, we have seen that many of the popular continuous time stochastic volatility models can be approximated by corresponding discrete

time stochastic volatility models. Similarly, one can understand a continuous time SV model as an approximation to corresponding discrete time SV models. One could further consider diffusion limits of specific given discrete time SV models after proper scaling, but we will not report on such results for stochastic volatility models, since the discrete time SV models obtained from continuous time SV models via (23) already cover a wide range of popular volatility models.

References

Barndorff-Nielsen, O.E. and Shephard, N. (2001a): Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, Series B* **63**, 167–241.

Barndorff-Nielsen, O.E. and Shephard, N. (2001b): Modelling by Lévy processes for financial econometrics. *In: Barndorff-Nielsen, O.E., Mikosch, T. and Resnick, S. (Eds.): Lévy Processes, Theory and Applications*, 283–318. Birkhäuser, Boston.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Brockwell, P.J. (2004): Representations of continuous time ARMA processes. *Journal of Applied Probability* **41A**, 375–382.

Brockwell, P. (2008): Lévy-driven continuous-time ARMA processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*,456–480. Springer, New York.

Brockwell, P.J., Erdenebaatar, C. and Lindner, A.M. (2006): Continuous time GARCH processes. *The Annals of Applied Probability* **16**, 790–826.

Brown, L.D., Wang, Y. and Zhao, L.H. (2003): On the statistical equivalence at suitable frequencies of GARCH and stochastic volatility models with the corresponding diffusion model. *Statistica Sinica* **13**, 993–1013.

Corradi, V. (2000) Reconsidering the continuous time limit of the GARCH(1,1) process *Journal of Econometrics*. **96**, 145–153.

Davis, R.A. and Mikosch, T. (2008): Probabilistic properties of stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 255–267. Springer, New York.

Drost, F.C. and Nijman, T.E. (1993): Temporal aggregation of GARCH processes. *Econometrica* **61**, 909–927.

Drost, F.C. and Werker, B.J.M. (1996): Closing the GARCH gap: Continuous time GARCH modeling. *Journal of Econometrics* **74**, 31–57.

Duan, J.-C. (1997): Augmented GARCH(p,q) process and its diffusion limit. *Journal of Econometrics* **79**, 97–127.

Engle, R.F. (1982): Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1008.

Fasen, V. (2008): Asymptotic results for sample autocovariance functions and extremes of integrated generalized Ornstein-Uhlenbeck processes. *Preprint.*

Ghysels, E., Harvey, A.C. and Renault, E. (1996): Stochastic volatility. *In: Maddala, G.S and Rao, C.R. (Eds.): Handbook of Statistics* **14**, 119–191. North-Holland, Amsterdam.

Haan, L. de and Karandikar, R.L. (1989): Embedding a stochastic difference equation into a continuous-time process. *Stochastic Processes and Their Applications* **32**, 213–224.

Haug, S. and Czado, C. (2007): An exponential continuous time GARCH process. *Journal of Applied Probability* **44**, 960–976.

Haug, S., Klüppelberg, C., Lindner, A. and Zapp, M. (2007): Method of moment estimation in the COGARCH(1,1) model. *The Econometrics Journal* **10**, 320–341.

Hull, J. and White, A. (1987): The pricing of options on assets with stochastic volatilities. *Journal of Finance* **42**, 281–300.

Jacod, J. and Shiryaev, A.N. (2003): *Limit Theorems for Stochastic Processes*. 2nd edition. Springer, Berlin Heidelberg New York.

Kallsen, J. and Taqqu, M.S. (1998): Option pricing in ARCH-type models. *Mathematical Finance* **8**, 13–26.

Kallsen, J. and Vesenmayer, B. (2008): COGARCH as a continuous-time limit of GARCH(1,1). *Stochastic Processes and Their Applications*, to appear.

Klüppelberg, C., Lindner, A. and Maller, R. (2004): A continuous-time GARCH process driven by a Lévy process: stationarity and second order behaviour. *Journal of Applied Probability* **41**, 601–622.

Klüppelberg, C., Lindner, A. and Maller, R. (2006): Continuous time volatility modelling: COGARCH versus Ornstein models. *In: Kabanov, Yu., Liptser, R. and Stoyanov, J. (Eds.): From Stochastic Calculus to Mathematical Finance. The Shiryaev Festschrift*, 393–419. Springer, Berlin Heidelberg New York.

Le Cam, L. (1986): *Asymptotic Methods in Statistical Decision Theory* Springer, Berlin Heidelberg New York.

Lindner, A.M. (2008): Stationarity, mixing, distributional properties and moments of GARCH(p, q). *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 43–69. Springer, New York.

Lorenz, R. (2006): Weak Approximation of Stochastic Delay Differential Equations with Bounded Memory by Discrete Time Series. *Ph.D. thesis*, Humboldt-Universität zu Berlin.

Maller, R.A., Müller, G. and Szimayer, A. (2008a): GARCH modelling in continuous time for irregularly spaced time series data. *Bernoulli* to appear.

Maller, R.A., Müller, G. and Szimayer, A. (2008b): Ornstein-Uhlenbeck processes and extensions. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 420–437. Springer, New York.

Meddahi, N. and Renault, E. (2004): Temporal aggregation of volatility models. *Journal of Econometrics* **19**, 355–379.

Nelson, D.B. (1990): ARCH models as diffusion approximations. *Journal of Econometrics* **45**, 7–38.

Protter, P.E. (2004) *Stochastic Integration and Differential Equations*. 2nd edition. Springer, Berlin Heidelberg New York.

Scott, L.O. (1987): Option pricing when the variance changes randomly: theory, estimation and an application. *Journal of Financial Quantitative Analysis* **22**, 419–439.

Shephard, N. and Andersen, T.G. (2008): Stochastic volatility: Origins and overview. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 233–254. Springer, New York.

Stelzer, R. (2008): Multivariate continuous time Lévy-driven GARCH processes. *Preprint*.

Szimayer, A. and Maller, R.A. (2007): Finite approximation schemes for Lévy processes, and their application to optimal stopping problems. *Stochastic Processes and Their Applications* **117**, 1422–1447.

Taylor, S.J. (1982): Financial returns modelled by the product of two stochastic processes – a study of daily sugar prices 1961-79. *In: Anderson, O.D. (Ed.): Time Series Analysis: Theory and Practice 1*, 203–226. North-Holland, Amsterdam.

Teräsvirta, T. (2008): An introduction to univariate GARCH models *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 17–42. Springer, New York.

Wang, Y. (2002): Asymptotic nonequivalence of GARCH models and diffusions. *The Annals of Statistics* **30**, 754–783.

Wiggins, J.B. (1987): Option values under stochastic volatility: theory and empirical estimates. *Journal of Financial Economics* **19**, 351–372.

Zheng, Z. (2005): Recrossing the bridge from discrete to continuous time towards a complete model with stochastic volatility I. *Preprint*.

Maximum Likelihood and Gaussian Estimation of Continuous Time Models in Finance

Peter C. B. Phillips and Jun Yu*

Abstract This paper overviews maximum likelihood and Gaussian methods of estimating continuous time models used in finance. Since the exact likelihood can be constructed only in special cases, much attention has been devoted to the development of methods designed to approximate the likelihood. These approaches range from crude Euler-type approximations and higher order stochastic Taylor series expansions to more complex polynomial-based expansions and infill approximations to the likelihood based on a continuous time data record. The methods are discussed, their properties are outlined and their relative finite sample performance compared in a simulation experiment with the nonlinear CIR diffusion model, which is popular in empirical finance. Bias correction methods are also considered and particular attention is given to jackknife and indirect inference estimators. The latter retains the good asymptotic properties of ML estimation while removing finite sample bias. This method demonstrates superior performance in finite samples.

Peter C. B. Phillips
Yale University, University of Auckland, University of York, and Singapore Management University, e-mail: peter.phillips@yale.edu

Jun Yu
School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, e-mail: yujun@smu.edu.sg

* Phillips gratefully acknowledges support from a Kelly Fellowship and the NSF under the Grant Nos. SES 04-142254 and SES 06-47086. Yu gratefully acknowledge financial support from the Ministry of Education AcRF Tier 2 fund under Grant No. T206B4301-RS. We wish to thank Yacine Aït-Sahalia and a referee for comments on an earlier version of the paper.

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_22,

1 Introduction

Continuous time models have provided a convenient mathematical framework for the development of financial economic theory (e.g., Merton (1990), asset pricing, and the modern field of mathematical finance that relies heavily on stochastic processes (Karatzas and Shreve (1991)). These models now dominate the option pricing literature, which has mushroomed over the last three decades from a single paper (Black and Scholes (1973)) to a vast subdiscipline with strong practical applications in the finance industry. Correspondingly, the econometric analysis of continuous time models has received a great deal attention in financial econometrics, providing a basis from which these models may be brought to data and used in practical applications. Much of the focus is on the econometric estimation of continuous time diffusion equations. Estimation not only provides parameter estimates which may be used directly in the pricing of financial assets and derivatives but also serves as a stage in the empirical analysis of specification and comparative diagnostics.

Many models that are used to describe financial time series are written in terms of a continuous time diffusion $X(t)$ that satisfies the stochastic differential equation

$$dX(t) = \mu(X(t);\theta)dt + \sigma(X(t);\theta)dB(t), \tag{1}$$

where $B(t)$ is a standard Brownian motion, $\sigma(X(t);\theta)$ is some specified diffusion function, $\mu(X(t);\theta)$ is a given drift function, and θ is a vector of unknown parameters. This class of parametric model has been widely used to characterize the temporal dynamics of financial variables, including stock prices, interest rates, and exchange rates.

It has been argued that when the model is correctly specified, the preferred choice of estimator and preferred basis for inference should be maximum likelihood (ML) – see, for example, Aït-Sahalia (2002) and Durham and Gallant (2002). Undoubtedly, the main justification for the use of the ML method lies in its desirable asymptotic properties, particularly its consistency and asymptotic efficiency under conditions of correct specification. In pursuit of this goal, various ML and Gaussian (that is, ML under Gaussian assumptions) methods have been proposed. Some of these methods involve discrete approximations, others are exact (or exact under certain limiting conditions on the approximation). Some are computationally inexpensive while others are computationally intensive. Some are limited to particular formulations, others have much wide applicability.

The purpose of the present chapter is to review this literature and overview the many different approaches to estimating continuous time models of the form given by (1) using ML and Gaussian methods. In the course of this overview, we shall discuss the existing methods of estimation and their merits and drawbacks. A simple Monte Carlo experiment is designed to reveal the finite sample performance of some of the most commonly used estima-

tion methods. The model chosen for the experiment is a simple example of (1) that involves a square root diffusion function. This model is popular in applied work for modeling short term interest rates and is known in the term structure literature as the Cox-Ingersoll-Ross or CIR model (see (9) below). One of the principal findings from this simulation experiment is that all ML methods, including "exact" methods, have serious *finite sample estimation bias* in the mean reversion parameter. This bias is significant even when the number of observations is as large as 500 or 1000. It is therefore important in ML/Gaussian estimation to take such bias effects into account. We therefore consider two estimation bias reduction techniques – the jackknife method and the indirect inference estimation – which may be used in conjunction with ML, Gaussian or various approximate ML methods. The indirect inference estimator demonstrates markedly superior results in terms of bias reduction and overall mean squared error in comparison with all other methods.

The chapter is organized as follows. Section 2 outlines the exact ML method, Section 3 and Section 4 review the literature on implementing ML/Gaussian methods in continuous time financial models and the practicalities of implementation. Section 5 reports a Monte Carlo study designed to investigate and compare the performance of some ML/Gaussian estimation methods for the CIR model. Section 6 reviews two bias reduction methods and examines their performance in the CIR model example. Section 7 briefly outlines some issues associated with extensions of ML/Gaussian procedures for multivariate models, and Section 8 concludes.

2 Exact ML Methods

2.1 ML based on the transition density

Assume the data $X(t)$ is recorded discretely at points $(h, 2h, \cdots, Nh(\equiv T))$ in the time interval $[0, T]$, where h is the discrete interval of observation of $X(t)$ and T is the time span of the data. The full sequence of N observations is $\{X_h, X_{2h}, \cdots, X_{Nh}\}$. If $X(t)$ is conceptualized for modeling purposes as annualized data which is observed discretely at monthly (weekly or daily) intervals, then $h = 1/12$ ($1/52$ or $1/252$). It is, of course, most convenient to assume that equi-spaced sampling observations are available and this assumption is most common in the literature, although it can be and sometimes is relaxed.

Many estimation methods are based on the construction of a likelihood function derived from the transition probability density of the discretely sampled data. This approach is explained as follows. Suppose $p(X_{ih}|X_{(i-1)h}, \theta)$ is the transition probability density. The Markov property of model (1) implies

the following log-likelihood function for the discrete sample[2]

$$\ell_{TD}(\theta) = \ln(p(X_{ih}|X_{(i-1)h}, \theta)). \tag{2}$$

The resulting estimator will be consistent, asymptotically normally distributed and asymptotically efficient under the usual regularity conditions for maximum likelihood estimation in (stationary) dynamic models (Hall and Heyde (1980); Billingsley (1961). In nonstationary, nonergodic cases, the limit theory is no longer asymptotically normal and there are several possibilities, including various unit root, local to unity, mildly explosive and explosive limit distributions (Phillips (1987), Chan and Wei (1988); Phillips (1991); Phillips and Magdalinos (2007).

To perform exact ML estimation, one needs a closed form expression for $\ell_{TD}(\theta)$ and hence $\ln(p(X_{ih}|X_{(i-1)h}, \theta))$. Unfortunately, only in rare cases, do the transition density and log likelihood component $\ln(p(X_{ih}|X_{(i-1)h}, \theta))$ have closed form analytical expressions. All other cases require numerical techniques or analytic or simulation-based approximants.

The following list reviews the continuous time models used in finance that have closed-form expressions for the transition density.

1. **Geometric Brownian Motion**:

$$dX(t) = \mu X(t)\, dt + \sigma X(t)\, dB(t). \tag{3}$$

Black and Scholes (1973) used this process to describe the movement of stock prices in their development of the stock option price formula. Since

$$d \ln X(t) = \frac{1}{X(t)} dX(t) - \frac{(dX(t))^2}{2X(t)^2} = \mu dt + \sigma dB(t) - \frac{1}{2}\sigma^2 dt, \tag{4}$$

the transformed process $\ln X(t)$ follows the linear diffusion

$$d \ln X(t) = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma\, dB(t). \tag{5}$$

As a result, $X_{ih}|X_{(i-1)h} \sim \text{LN}((\mu - \frac{\sigma^2}{2})h + \ln(X_{(i-1)h}), \sigma^2 h)$, where LN denotes the log-normal distribution.

2. **Ornstein-Uhlenbeck (OU) process (or Vasicek model)** :

$$dX(t) = \kappa(\mu - X(t))dt + \sigma\, dB(t). \tag{6}$$

Vasicek (1977) used this process to describe the movement of short term interest rates. Phillips (1972) showed that the exact discrete model cor-

[2] Our focus in the present discussion is on the usefulness of the transition density for estimation purposes. But we note that the transition density is needed and used for many other applications, such as for pricing derivatives and for obtaining interval and density forecasts.

responding to (6) is given by

$$X_{ih} = e^{-\kappa h} X_{(i-1)h} + \mu\left(1 - e^{-\kappa h}\right) + \sigma\sqrt{(1 - e^{-2\kappa h})/(2\kappa)}\epsilon_i, \quad (7)$$

where $\epsilon_i \sim$ i.i.d. $N(0, 1)$. Phillips (1972) also developed an asymptotic theory for nonlinear least squares/ML estimates of the parameters in a multivariate version of (6) using the exact discrete time model (7), showing consistency, asymptotic normality and efficiency under stationarity assumptions ($\kappa > 0$ in the univariate case here). The transition density for the Vasicek model follows directly from (7) and is

$$X_{ih}|X_{(i-1)h} \sim N\left(\mu(1 - e^{-\kappa h}) + e^{-\kappa h} X_{(i-1)h}, \sigma^2(1 - e^{-2\kappa h})/(2\kappa)\right). \quad (8)$$

3. **Square-root (or Cox-Ingersoll-Ross) model** :

$$dX(t) = \kappa(\mu - X(t))dt + \sigma\sqrt{X(t)}\, dB(t). \quad (9)$$

Cox, Ingersoll and Ross (1985), CIR hereafter, also used this process to describe movements in short term interest rates. The exact discrete model corresponding to (9) is given by

$$X_{ih} = e^{-\kappa h} X_{(i-1)h} + \mu\left(1 - e^{-\kappa h}\right) + \sigma \int_{(i-1)h}^{ih} e^{-\kappa(ih-s)}\sqrt{X(s)} dB\left(s\right). \quad (10)$$

When $2\kappa\mu/\sigma^2 \geq 1$, X is distributed over the positive half line. Feller (1952) showed that the transition density of the square root model is given by

$$X_{ih}|X_{(i-1)h} = ce^{-u-v}(v/u)^{q/2} I_q(2(uv)^{1/2}) \quad (11)$$

where $c = 2\kappa/(\sigma^2(1-e^{-\kappa h}))$, $u = cX_{(i-1)h}e^{-\kappa h}$, $v = cX_{ih}$, $q = 2\kappa\mu/\sigma^2 - 1$, and $I_q(\cdot)$ is the modified Bessel function of the first kind of order q.

4. **Inverse square-root model** :

$$dX(t) = \kappa(\mu - X(t))X(t)dt + \sigma X^{1.5}(t)\, dB(t). \quad (12)$$

Ahn and Gao (1999) again used this process to model short term interest rates. When $\kappa, \mu > 0$, X is distributed over the positive half line. Ahn and Gao (1999) derived the transition density of the inverse square root model as

$$X_{ih}|X_{(i-1)h} = c^{-1}e^{-u-v}(v)^{q/2+2}u^{-q/2} I_q(2(uv)^{1/2}) \quad (13)$$

where $c = 2\kappa\mu/(\sigma^2(1 - e^{-\kappa\mu h}))$, $u = ce^{-\kappa\mu h}/X_{(i-1)h}$, $v = c/X_{ih}$, $q = 2(\kappa + \sigma^2)/\sigma^2 - 1$.

2.2 ML based on the continuous record likelihood

If a continuous sample path of the process $X(t)$ was recorded over the interval $[0, T]$, direct ML estimation would be possible based on the continuous path likelihood. This likelihood is very useful in providing a basis for the so-called continuous record or infill likelihood function and infill asymptotics in which a discrete record becomes continuous by a process of infilling as the sampling interval $h \to 0$. Some of these infill techniques based on the continuous record likelihood are discussed later in Section 4. Since financial data are now being collected on a second by second and tick by tick basis, this construction is becoming much more important.

When $X(t)$ is observed continuously, a log-likelihood function for the continuous record $\{X(t)\}_{t=0}^{T}$ may be obtained directly from the Radon Nikodym (RN) derivative of the relevant probability measures. The RN derivative produces the relevant probability density and can be regarded as a change of measure among the absolutely continuous probability measures, the calculation being facilitated by the Girsanov theorem (e.g., Karatzas and Shreve (1991)). The approach is convenient and applies quite generally to continuous time models with flexible drift and diffusion functions.

In the stochastic process literature the quadratic variation or square bracket process is well known to play an important role in the study of stochastic differential equations. In the case of equation (1), the square bracket process of $X(t)$ has the explicit form

$$[X]_T = \int_0^T (dX(t))^2 = \int_0^T \sigma^2(X(t);\theta)dt, \tag{14}$$

which is a continuously differentiable increasing function. In fact, we have $d[X]_t = \sigma(X(t);\theta)^2 dt$. In consequence, when a continuous sample path of the process $X(t)$ is available, the quadratic variation of X provides a perfect estimate of the diffusion function and hence the parameters on which it depends, provided these are identifiable in $\sigma^2(X(t);\theta)$. Thus, with the availability of a continuous record, we can effectively assume the diffusion term (i.e., $\sigma(X(t);\theta) = \sigma(X(t))$ is known and so this component does not involve any unknown parameters. It follows that the exact continuous record or infill log-likelihood can be constructed via the Girsanov theorem (e.g., Liptser and Shiryaev (2000)) as

$$\ell_{IF}(\theta) = \int_0^T \frac{\mu(X(t);\theta)}{\sigma^2(X(t))} dX(t) - \frac{1}{2}\int_0^T \frac{\mu^2(X(t);\theta)}{\sigma^2(X(t))} dt. \tag{15}$$

In this likelihood, the parameter θ enters via the drift function $\mu(X(t);\theta)$. Lánska (1979) established the consistency and asymptotic normality of the continuous record ML estimator of θ when $T \to \infty$ under certain regularity conditions.

To illustrate the approach, consider the following OU process,

$$dX(t) = \kappa X(t)dt + \sigma_0 dB(t), \tag{16}$$

where σ_0 is known and κ is the only unknown parameter. The exact log-likelihood in this case is given by

$$\ell_{IF}(\kappa) = \int_0^T \frac{\kappa X(t)}{\sigma_0^2} dX(t) - \frac{1}{2}\int_0^T \frac{\kappa^2 X^2(t)}{\sigma_0^2} dt, \tag{17}$$

and maximizing the log-likelihood function immediately gives rise to the following ML estimator of κ:

$$\widehat{\kappa} = \left(\int_0^T X^2(t)dt\right)^{-1} \int_0^T X(t)dX(t) \tag{18}$$

This estimator is analogous in form to the ML/OLS estimator of the autoregressive coefficient in the discrete time Gaussian autoregression

$$X_t = \phi X_{t-1} + \epsilon_t,\ \epsilon_t \sim \text{i.i.d. } N(0,1) \tag{19}$$

viz., $\widehat{\phi} = \left(\sum_{t=1}^n X_{t-1}^2\right)^{-1} \sum_{t=1}^n X_t X_{t-1}$. It is also interesting to observe that when $\kappa = 0$ (18) has the same form as the limit distribution of the (discrete time) autoregressive coefficient estimator when $\phi = 1$ in (19). These connections with unit root limit theory are explored in Phillips (1987).

In practice, of course, a continuous record of $\{X(t)\}_{t=0}^T$ is not available and estimators such as (18) are infeasible. On the other hand, as the sampling interval h shrinks, discrete data may be used to produce increasingly good approximations to the quadratic variation (14), the continuous record likelihood (15) and estimators such as (18). These procedures may be interpreted as infill likelihood methods in that they replicate continuous record methods by infilling the sample record as $h \to 0$.

3 Approximate ML Methods Based on Transition Densities

Except for a few special cases such as those discussed earlier, the transition density does not have a closed-form analytic expression. As a result, the exact ML method discussed in Section 2.1 is not generally applicable. To address this complication, many alternative approaches have been developed. The methods involve approximating the transition densities, the model itself or the likelihood function. This section reviews these methods.

3.1 The Euler approximation and refinements

The Euler scheme approximates a general diffusion process such as equation (1) by the following discrete time model

$$X_{ih} = X_{(i-1)h} + \mu(X_{(i-1)h}, \theta)h + \sigma(X_{(i-1)h}, \theta)\sqrt{h}\epsilon_i, \tag{20}$$

where $\epsilon_i \sim$ i.i.d. $N(0,1)$. The transition density for the Euler discrete time model has the following closed form expression:

$$X_{ih}|X_{(i-1)h} \sim N\left(X_{(i-1)h} + \mu(X_{(i-1)h}, \theta)h, \sigma^2(X_{(i-1)h}, \theta)h\right). \tag{21}$$

For the Vasicek model, the Euler discrete approximation is of the form

$$X_{ih} = \kappa\mu h + (1 - \kappa h)X_{(i-1)h} + \sigma N(0, h). \tag{22}$$

Comparing the approximation (22) with the exact discrete time model (7), we see that $\kappa\mu h$, $1 - \kappa h$ and $\sigma^2 h$ replace $\mu(1 - e^{-\kappa h})$, $e^{-\kappa h}$, and $\sigma^2(1 - e^{-2\kappa h})/(2\kappa)$, respectively. These replacements may be motivated by considering the first order term in the following Taylor expansions:

$$\mu(1 - e^{-\kappa h}) = \kappa\mu h + O(h^2), \tag{23}$$

$$e^{-\kappa h} = 1 - \kappa h + O(h^2), \tag{24}$$

$$\sigma^2(1 - e^{-2\kappa h})/(2\kappa) = \sigma^2 h + O(h^2). \tag{25}$$

Obviously, when h is small, the Euler scheme should provide a good approximation to the exact discrete time model. However, when h is large, the Euler approximation can be poor. To illustrate magnitude of the approximation error, first consider the case where $\kappa = 1$ and $h = 1/12$, in which case $e^{-\kappa h}$ is 0.92 whereas $1 - \kappa h$ is 0.9167 and the approximation is good. But if $\kappa = 1$ and $h = 1$, then $e^{-\kappa h}$ is 0.3679 whereas $1 - \kappa h$ is 0. These comparisons suggest that the Euler discretization offers a good approximation to the exact discrete time model for daily or higher frequencies but not for annual or lower frequencies. The bias introduced by this discrete time approximation is called the *discretization bias.*

The advantages of the Euler method include the ease with which the likelihood function is obtained, the low computational cost, and the wide range of its applicability. The biggest problem with the procedure is that when h is fixed the estimator is inconsistent (Merton (1980); Lo (1988)). The magnitude of the inconsistency can be analyzed, using the methods of Sargan (1974), in terms of the observation interval h. Lo (1988) illustrated the size of inconsistency in the context of model (3).

A closely related discretization method, suggested by Bergstrom (1966) and Houthakker and Taylor (1966), is based on integrating the stochastic differential equation and using the following trapezoidal rule approximation

$$\int_{(i-1)h}^{ih} \mu(X(t);\theta)dt = \frac{h}{2}\left\{\mu(X_{ih};\theta) + \mu(X_{(i-1)h};\theta)\right\}. \tag{26}$$

For the OU process the corresponding discrete approximate model is given by

$$X_{ih} - X_{(i-1)h} = \kappa\mu - \frac{\kappa h}{2}\left(X_{ih} + X_{(i-1)h}\right) + \sigma N(0,h), \tag{27}$$

which involves the current period observation X_{ih} on both sides of the equation. Solving (27) we obtain

$$\begin{aligned} X_{ih} &= \frac{\kappa\mu h}{\left(1+\frac{\kappa h}{2}\right)} + \frac{1-\frac{\kappa h}{2}}{1+\frac{\kappa h}{2}}X_{(i-1)h} + \frac{\sigma}{\left(1+\frac{\kappa h}{2}\right)}N(0,h) \\ &= \kappa\mu h + (1-\kappa h)X_{(i-1)h} + \sigma N(0,h) + O\left(h^{3/2}\right), \end{aligned}$$

so that the Bergstrom approximation is equivalent to the Euler approximation to $O(h)$. In the multivariate case, the Bergstrom approximation leads to a non-recursive simultaneous equations model approximation to a system of recursive stochastic differential equations. The resulting system may be estimated by a variety of simultaneous equations estimators, such as instrumental variables, for example by using lagged X values as instruments. Again, the magnitude of the inconsistency may be analyzed in terms of the observation interval h, as in Sargan (1974) who showed the asymptotic bias in the estimates to be typically of $O\left(h^2\right)$.

There are a number of ways to reduce the discretization bias induced by the Euler approximation. Before we review these refinements, it is important to emphasize that the aim of these refinements is simply bias reduction.

Elerian (1998) suggests using the scheme proposed by Milstein (1978). The idea is to take a second order term in a stochastic Taylor series expansion to refine the Euler approximation (20). We proceed as follows. Integrating (1) we have

$$\int_{(i-1)h}^{ih} dX(t) = \int_{(i-1)h}^{ih} \mu(X(t);\theta)dt + \int_{(i-1)h}^{ih} \sigma(X(t);\theta)dB(t), \tag{28}$$

and by stochastic differentiation we have

$$\begin{aligned} d\mu(X(t);\theta) &= \mu'(X(t);\theta)dX(t) + \frac{1}{2}\mu''(X(t);\theta)\left(dX(t)\right)^2 \\ &= \mu'(X(t);\theta)dX(t) + \frac{1}{2}\mu''(X(t);\theta)\sigma^2(X(t);\theta)dt, \end{aligned}$$

and

$$d\sigma(X(t);\theta) = \sigma'(X(t);\theta)dX\,(t) + \frac{1}{2}\sigma''(X(t);\theta)\sigma^2(X(t);\theta)dt, \tag{29}$$

so that

$$\begin{aligned}\mu(X\,(t)\,;\theta) &= \mu(X_{(i-1)h};\theta) + \int_{(i-1)h}^{t} \mu'(X(s);\theta)dX\,(s) \\ &\quad + \frac{1}{2}\int_{(i-1)h}^{t} \mu''(X(s);\theta)\sigma^2(X(s);\theta)ds \\ &= \mu(X_{(i-1)h};\theta) + \int_{(i-1)h}^{t} \mu'(X(s);\theta)\mu(X(s);\theta)ds + \\ &\quad \frac{1}{2}\int_{(i-1)h}^{t} \mu''(X(s);\theta)\sigma^2(X(s);\theta)ds + \\ &\quad \int_{(i-1)h}^{t} \mu'(X(s);\theta)\sigma(X(s);\theta)dB(s),\end{aligned}$$

and

$$\begin{aligned}\sigma(X\,(t)\,;\theta) &= \sigma(X_{(i-1)h};\theta) + \int_{(i-1)h}^{t} \sigma'(X(s);\theta)\mu(X(s);\theta)ds + \\ &\quad \frac{1}{2}\int_{(i-1)h}^{t} \sigma''(X(s);\theta)\sigma^2(X(s);\theta)ds + \\ &\quad \int_{(i-1)h}^{t} \sigma'(X(s);\theta)\sigma(X(s);\theta)dB(s),\end{aligned}$$

with $\sigma'(X_{(i-1)h};\theta) = [\partial\sigma(X;\theta)/\partial X]_{X=X_{(i-1)h}}$. Substituting these expressions into (28) we obtain

$$\begin{aligned}X_{ih} - X_{(i-1)h} &= \mu(X_{(i-1)h};\theta)h + \sigma(X_{(i-1)h};\theta)\int_{(i-1)h}^{ih} dB\,(t) \\ &\quad + \int_{(i-1)h}^{ih}\int_{(i-1)h}^{t} \sigma'(X(s);\theta)\sigma(X(s);\theta)dB(s)dB\,(t) + R,\end{aligned} \tag{30}$$

where R is a remainder of smaller order. Upon further use of the Itô formula on the penultimate term of (31), we obtain the following refinement of the Euler approximation

$$X_{ih} - X_{(i-1)h} \simeq \mu(X_{(i-1)h};\theta)h + \sigma(X_{(i-1)h};\theta)\int_{(i-1)h}^{ih} dB\,(t) + \sigma'(X_{(i-1)h};\theta)\sigma(X_{(i-1)h};\theta)\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t} dB(s)dB\,(t)\,,$$

The multiple stochastic integral has the following reduction

$$\begin{aligned}
&\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t} dB(s)dB\,(t) \\
&= \int_{(i-1)h}^{ih} \left(B(t) - B_{(i-1)h}\right) dB\,(t) \\
&= \int_{(i-1)h}^{ih} B(t)dB\,(t) - B_{(i-1)h}\left(B_{ih} - B_{(i-1)h}\right) \\
&= \frac{1}{2}\left\{\left(B_{ih}^2 - B_{(i-1)h}^2\right) - h\right\} - B_{(i-1)h}\left(B_{ih} - B_{(i-1)h}\right) \\
&= \frac{1}{2}\left\{\left(B_{ih} - B_{(i-1)h}\right)^2 - h\right\},
\end{aligned}$$

Then the refined Euler approximation can be written as

$$\begin{aligned}
X_{ih} - X_{(i-1)h} &\simeq \mu(X_{(i-1)h};\theta)h + \sigma(X_{(i-1)h};\theta)\left(B_{ih} - B_{(i-1)h}\right) \\
&\quad + \sigma'(X_{(i-1)h};\theta)\sigma(X_{(i-1)h};\theta)\frac{1}{2}\left\{\left(B_{ih} - B_{(i-1)h}\right)^2 - h\right\} \\
&= \left\{\mu(X_{(i-1)h};\theta) - \frac{1}{2}\sigma'(X_{(i-1)h};\theta)\sigma(X_{(i-1)h};\theta)\right\}h \\
&\quad + \sigma(X_{(i-1)h};\theta)\left(B_{ih} - B_{(i-1)h}\right) \\
&\quad + \frac{1}{2}\sigma'(X_{(i-1)h};\theta)\sigma(X_{(i-1)h};\theta)\left(B_{ih} - B_{(i-1)h}\right)^2
\end{aligned}$$

The approach to such refinements is now very well developed in the numerical analysis literature and higher order developments are possible - see Kloeden and Platen (1999) for an extensive review.

It is convenient to write $B_{ih} - B_{(i-1)h} = \sqrt{h}\epsilon_i$ where ϵ_i is standard Gaussian. Then, the Milstein approximation to model (1) produces the following discrete time model:

$$\begin{aligned}
X_{ih} = X_{(i-1)h} &+ \mu(X_{(i-1)h},\theta)h - g(X_{(i-1)h},\theta)h \\
&+ \sigma(X_{(i-1)h},\theta)\sqrt{h}\epsilon_i + g(X_{(i-1)h},\theta)h\epsilon_i^2,
\end{aligned} \tag{31}$$

where

$$g(X_{(i-1)h},\theta) = \frac{1}{2}\sigma'(X_{(i-1)h};\theta)\sigma(X_{(i-1)h};\theta). \tag{32}$$

While Elerian (1998) used the Milstein scheme in connection with a simulation based approach, Tse, Zhang and Yu (2004) used the Milstein scheme in a Bayesian context. Both papers document some improvement from the Milstein scheme over the Euler scheme.

Kessler (1997) advocated approximating the transition density using a Gaussian density whose conditional mean and variance are obtained using higher order Taylor expansions. For example, the second-order approximation leads to the following discrete time model:

$$X_{ih} = \widehat{\mu}(X_{(i-1)h};\theta) + \widehat{\sigma}(X_{(i-1)h};\theta)\epsilon_i, \tag{33}$$

where

$$\widehat{\mu}(X_{(i-1)h};\theta) = X_{(i-1)h} + \mu(X_{(i-1)h};\theta)h + \left(\mu(X_{(i-1)h};\theta)\mu'(X_{(i-1)h};\theta) + \frac{\sigma^2(X_{(i-1)h};\theta)\mu''(X_{(i-1)h};\theta)}{2}\right)\frac{h}{2}$$

and

$$\begin{aligned}\widehat{\sigma}^2(X_{(i-1)h};\theta) &= X_{(i-1)h}^2 + \left(2\mu(X_{(i-1)h};\theta)X_{(i-1)h} + \sigma^2(X_{(i-1)h};\theta)\right)h \\ &= \{2\mu(X_{(i-1)h};\theta)(2\mu'(X_{(i-1)h};\theta)X_{(i-1)h} + \mu(X_{(i-1)h};\theta) \\ &\quad +\sigma(X_{(i-1)h};\theta)\sigma'(X_{(i-1)h};\theta)) + \sigma^2(X_{(i-1)h};\theta)\times \\ &\quad [\mu''(X_{(i-1)h};\theta)X_{(i-1)h} + 2\mu(X_{(i-1)h};\theta) + (\sigma'(X_{(i-1)h};\theta))^2 \\ &\quad +\sigma(X_{(i-1)h};\theta)\sigma'(X_{(i-1)h};\theta)]\}\frac{h^2}{2} - \widehat{\mu}^2(X_{(i-1)h};\theta).\end{aligned}$$

Nowman (1997) suggested an approach which assumes that the conditional volatility remains unchanged over the unit intervals, $[(i-1)h, ih)$, $i = 1, 2..., N$. In particular, he approximates the model:

$$dX(t) = \kappa(\mu - X(t))dt + \sigma(X(t),\theta)dB(t) \tag{34}$$

by

$$dX(t) = \kappa(\mu - X(t))dt + \sigma(X_{(i-1)h};\theta)dB(t), \quad (i-1)h \le t < ih. \tag{35}$$

It is known from Phillips (1972) and Bergstrom (1984) that the exact discrete model of (35) has the form

$$X_{ih} = e^{-\kappa h}X_{(i-1)h} + \mu\left(1 - e^{-\kappa h}\right) + \sigma(X_{(i-1)h};\theta)\sqrt{\frac{1-e^{-2\kappa h}}{2\kappa}}\epsilon_i, \tag{36}$$

where $\epsilon_i \sim$ i.i.d. $N(0,1)$. With this approximation, the Gaussian ML method can be used to estimate equation (36) directly. This method also extends in a straightforward way to multivariate systems. The Nowman procedure can

be understood as applying the Euler scheme to the diffusion term over the unit interval. Compared with the Euler scheme where the approximation is introduced to both the drift function and the diffusion function, the Nowman method can be expected to reduce some of the discretization bias, as the treatment of the drift term does not involve an approximation at least in systems with linear drift.

Nowman's method is related to the local linearization method proposed by Shoji and Ozaki (1997, 1998) for estimating diffusion processes with a constant diffusion function and a possible nonlinear drift function, that is

$$dX(t) = \mu(X(t); \theta)dt + \sigma dB(t). \tag{37}$$

While Nowman approximates the nonlinear diffusion term by a locally linear function, Shoji and Ozaki (1998) approximate the drift term by a locally linear function. The local linearization method can be used to estimate a diffusion process with a nonlinear diffusion function, provided that the process can be first transformed to make the diffusion function constant. This is achieved by the so-called Lamperti transform which will be explained in detailed below.

While all these refinements offer some improvements over the Euler method, with a fixed h, all the estimators remain inconsistent. As indicated, the magnitude of the inconsistency or bias may analyzed in terms of its order of magnitude as $h \to 0$. This appears only to have been done by Sargan (1974), Phillips (1974) and Lo (1988) for linear systems and some special cases.

3.2 Closed–form approximations

The approaches reviewed above seek to approximate continuous time models by discrete time models, the accuracy of the approximations depending on the sampling interval h. Alternatively, one can use closed-form sequences to approximate the transition density itself, thereby developing an approximation to the likelihood function. Two different approximation mechanisms have been proposed in the literature. One is based on Hermite polynomial expansions whereas the other is based on the saddlepoint approximation.

3.2.1 Hermite expansions

This approach was developed in Aït-Sahalia (2002) and illustrated in Aït-Sahalia (1999). Before obtaining the closed-form expansions, a Lamperti transform (mentioned earlier) is performed on the continuous time model so that the diffusion function becomes a constant. The transformation has the

form $Y(t) = G(X(t))$, where $G'(x) = 1/\sigma(x;\cdot)$. The transformation is variance stabilizing and leads to another diffusion $Y(t)$, which by Itô's lemma can be shown to satisfy the stochastic differential equation

$$dY(t) = \mu_Y(Y(t);\theta)dt + dB(t), \tag{38}$$

where

$$\mu_Y(Y(t);\theta) = \frac{\mu(G^{-1}(Y);\theta)}{\sigma(G^{-1}(Y);\theta)} - \frac{1}{2}\sigma'(G^{-1}(Y);\theta). \tag{39}$$

A Hermite polynomial expansion of the transition density $p(Y_{ih}|Y_{(i-1)h},\theta)$ around the normal distribution leads to

$$p(Y_{ih}|Y_{(i-1)h},\theta) \approx h^{-1/2}\phi\left(\frac{Y_{ih}-Y_{(i-1)h}}{h^{1/2}}\right)\exp\left(\int_{Y_{(i-1)h}}^{Y_{ih}}\mu_Y(\omega;\theta)d\omega\right)\times \sum_{k=0}^{K} c_k(Y_{ih}|Y_{(i-1)h};\theta)\frac{h^k}{k!}, \tag{40}$$

where $\phi(\cdot)$ is the standard normal density function, $c_0(Y_{ih}|Y_{(i-1)h}) = 1$,

$$c_j(Y_{ih}|Y_{(i-1)h}) = j(Y_{ih}-Y_{(i-1)h})^{-j}\int_{Y_{(i-1)h}}^{Y_{ih}}(\omega - Y_{(i-1)h})^{j-1}\times \{\lambda_{Y_{ih}}(\omega;\theta)c_{j-1}(\omega|Y_{(i-1)h};\theta) + \frac{1}{2}\partial^2 c_{j-1}(\omega|Y_{(i-1)h};\theta)/\partial\omega^2\}d\omega,$$

$\forall j \geq 1$ and

$$\lambda_Y(y;\theta) = -\frac{1}{2}\left(\mu_Y^2(y;\theta) + \partial\mu_Y(y;\theta)/\partial y\right). \tag{41}$$

Under some regular conditions, Aït-Sahalia (2002) showed that when $K \to \infty$, the Hermite expansions (i.e., the right hand right in Equation (40)) approaches the true transition density. When applied to various interest rate models, Aït-Sahalia (1999) has found negligible approximation errors even for small values of K. Another advantage of this approach is that it is in closed-form and hence numerically tractable.

The approach described above requires the Lamperti transform be feasible. Aït-Sahalia (2007) and Bakshi and Ju (2005) proposed some techniques which avoid the Lamperti transform. Furthermore, Aït-Sahalia and Kimmel (2005, 2007) discussed how to use the method to estimate some latent variable models.

3.2.2 Saddlepoint approximations

The leading term in the Hermite expansions is normal whose tails may be too thin and the shape too symmetric relative to the true transition density. When this is the case, a moderately large value of K may be needed to ensure a good approximation of the Hermite expansion. An alternative approach is to choose a better approximating distribution as the leading term. One way to achieve this is to use a saddlepoint approximation.

The idea of the saddlepoint approximations is to approximate the conditional cumulant generating function of the transition density by means of a suitable expansion, followed by a careful choice of integration path in the integral that defines the transition density so that most of the contribution to the integral comes from integrating in the immediate neighborhood of a saddlepoint. The method was originally explored in statistics by Daniels (1953). Phillips (1978) developed a saddlepoint approximation to the distribution of ML estimator of the coefficient in discrete time first order autoregression, while Holly and Phillips (1979) proposed saddlepoint approximations for the distributions of k-class estimators of structural coefficients in simultaneous equation systems. There has since been a great deal of interest in the method in statistics - see Reid (1988), Field and Ronchetti (1990) and Butler (2007) for partial overviews of the field. Aït-Sahalia and Yu (2006) proposed the use of saddlepoint approximations to the transition density of continuous time models, which we now consider.

Let $\varphi_{X_{(i-1)h}}(u;\theta)$ be the conditional characteristic function corresponding to the transition density, viz.,

$$\varphi_{X_{(i-1)h}}(u;\theta) = E[\exp(uX_{ih}|X_{(i-1)h})]. \tag{42}$$

The conditional cumulant generating function is

$$K_{X_{(i-1)h}}(u;\theta) = \ln(\varphi_{X_{(i-1)h}}(u;\theta)). \tag{43}$$

The transition density has the following integral representation by Fourier inversion:

$$\begin{aligned} p(X_{ih}|X_{(i-1)h},\theta) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp(-iX_{ih}u)\varphi_{X_{(i-1)h}}(iu;\theta)du \\ &= \frac{1}{2\pi}\int_{\widehat{u}-i\infty}^{\widehat{u}+i\infty} \exp(-uX_{ih})\varphi_{X_{(i-1)h}}(u;\theta)du \\ &= \frac{1}{2\pi}\int_{\widehat{u}-i\infty}^{\widehat{u}+i\infty} \exp(K_{X_{(i-1)h}}(u;\theta) - uX_{ih})du \end{aligned} \tag{44}$$

Applying a Taylor expansion to $K_{X_{(i-1)h}}(u;\theta) - uX_{ih}$ around the saddlepoint $\widehat{u}$, one gets

$$K_{X_{(i-1)h}}(u;\theta) - uX_{ih} = K_{X_{(i-1)h}}(\widehat{u};\theta) - \widehat{u}X_{ih} - \frac{1}{2}\frac{\partial^2 K_{X_{(i-1)h}}(\widehat{u};\theta)}{\partial u^2}\nu - \frac{1}{6}\frac{\partial^3 K_{X_{(i-1)h}}(\widehat{u};\theta)}{\partial u^3}i\nu^3 + O(\nu^4).$$

Substituting this expansion to (43), one obtains a saddlepoint approximation to the integral, which involves the single leading term of the form

$$\frac{\exp(K_{X_{(i-1)h}}(\widehat{u};\theta) - uX_{ih})}{\sqrt{2\pi}\left(\frac{\partial^2 K_{X_{(i-1)h}}(\widehat{u};\theta)}{\partial u^2}\right)^{1/2}}, \tag{45}$$

and higher order terms of small order. As shown in Daniels (1954), the method has the advantage of producing a smaller relative error than Edgeworth and Hermite expansions.

When applying this method to transition densities for some continuous time models that are widely used in finance, Aït-Sahalia and Yu (2006) have found very small approximation errors. The method requires the saddlepoint to be analytically available or at least numerically calculable, an approach considered in Phillips (1984) that widens the arena of potential application. The saddlepoint method also requires the moment generating function of the transition density to exist, so that all moments of the distribution must be finite and heavy tailed transition distributions are therefore excluded. Multivariate extensions are possible using extensions of the saddlepoint method to this case - see Phillips (1980,1984), Tierney and Kadane (1986) and McCullagh (1987).

3.3 Simulated infill ML methods

As explained above, the Euler scheme introduces discretization bias. The magnitude of the bias is determined by h. When the sampling interval is arbitrarily small, the bias becomes negligible. One way of making the sampling interval arbitrarily small is to partition the original interval, say $[(i-1)h, ih]$, so that the new subintervals are sufficiently fine for the discretization bias to be negligible. By making the subintervals smaller, one inevitably introduces latent (that is, unobserved) variables between $X_{(i-1)h}$ and X_{ih}. To obtain the required transition density $p(X_{ih}|X_{(i-1)h}, \theta)$, these latent observations must be integrated out. When the partition becomes finer, the discretization bias is closer to 0 but the required integration becomes high dimensional. We call this approach to bias reduction the simulated infill ML method.

To fix ideas, suppose $M-1$ auxiliary points are introduced between $(i-1)h$ and ih, i.e.,

$$((i-1)h \equiv)\tau_0, \tau_1, \cdots, \tau_{M-1}, \tau_M(\equiv ih). \tag{46}$$

The Markov property implies that

$$p(X_{ih}|X_{(i-1)h};\theta) = \int \cdots \int p(X_{\tau_M}, X_{\tau_{M-1}}, \cdots, X_{\tau_1}|X_{\tau_0};\theta) dX_{\tau_1} \cdots dX_{\tau_{M-1}}$$
$$= \int \cdots \int \prod_{m=1}^{M} p(X_{\tau_m}|X_{\tau_{m-1}};\theta) dX_{\tau_1} \cdots dX_{\tau_{M-1}}. \quad (47)$$

The idea behind the simulated infill ML method is to approximate the densities $p(X_{\tau_m}|X_{\tau_{m-1}};\theta)$ (step 1) and then evaluate the multidimensional integral using importance sampling techniques (step 2). Among the class of simulated infill ML methods that have been suggested, Pedersen (1995) is one of the earliest contributions.

Pedersen suggested approximating the latent transition densities $p(X_{\tau_m}|X_{\tau_{m-1}};\theta)$ based on the Euler scheme and approximating the integral by drawing samples of $(X_{\tau_{M-1}}, \cdots, X_{\tau_1})$ via simulations from the Euler scheme. That is, the importance sampling function is the mapping from $(\epsilon_1, \epsilon_2, \cdots, \epsilon_{M-1}) \mapsto (X_{\tau_1}, X_{\tau_2}, \cdots, X_{\tau_{M-1}})$ given by the Euler scheme:

$$X_{\tau_{m+1}} = X_{\tau_m} + \mu(X_{\tau_m};\theta)h/M + \sigma(X_{\tau_m},\theta)\sqrt{h/M}\epsilon_{m+1},\ m = 0, \cdots, M-2, \quad (48)$$

where $(\epsilon_1, \epsilon_2, \cdots, \epsilon_{M-1})$ is a multivariate standard normal.

As noted in Durham and Gallant (2002), there are two sources of approximation error in Pedersen's method. One is the (albeit reduced) discretization bias in the Euler scheme. The second is due to the Monte Carlo integration. These two errors can be further reduced by increasing the number of latent infill points and the number of simulated paths, respectively. However, the corresponding computational cost will inevitably be higher.

In order to reduce the discretization bias in step 1, Elerian (1998) suggested replacing the Euler scheme with the Milstein scheme while Durham and Gallant advocated using a variance stablization transformation, i.e., applying the Lamperti transform to the continuous time model. Certainly, any method that reduces the discretization bias can be used. Regarding step 2, Elerian et al (2001) argued that the importance sampling function of Pedersen ignores the end-point information, X_{τ_M}, and Durham and Gallant (2002) showed that Pedersen's importance function draws most samples from regions where the integrand has little mass. Consequently, Pedersen's method is simulation-inefficient.

To improve the efficiency of the importance sampler, Durham and Gallant (2002) considered the following importance sampling function

$$X_{\tau_{m+1}} = X_{\tau_m} + \frac{X_{ih} - X_{\tau_m}}{ih - \tau_m} h/M + \sigma(X_{\tau_m},\theta)\sqrt{h/M}\epsilon_{m+1},\ m = 0, \cdots, M-2, \quad (49)$$

where $(\epsilon_1, \epsilon_2, \cdots, \epsilon_{M-1})$ is a multivariate standard normal. Loosing speaking, this is a Brownian bridge because it starts from $X_{(i-1)h}$ at $(i-1)h$ and is conditioned to terminate with X_{ih} at ih.

Another importance sampling function proposed by Durham and Gallant (2002) is to draw $X_{\tau_{m+1}}$ from the density $N(X_{\tau_m} + \widetilde{\mu}_m h/M, \widetilde{\sigma}_m^2 h/M)$ where $\widetilde{\mu}_m = (X_{\tau_M} - X_{\tau_m})/(ih - \tau_m)$, $\widetilde{\sigma}_m^2 = \sigma^2(X_{\tau_m})(M-m-1)/(M-m)$.

Elerian et al. (2001) proposed a more efficient importance function which is based on the following tied-down process:

$$p(X_{\tau_1}, \cdots, X_{\tau_{M-1}} | X_{\tau_0}, X_{\tau_M}). \tag{50}$$

In particular, they proposed using the Laplace approximation (c.f., Phillips (1984); Tierney and Kadane (1986)) to the tied-down process. That is, they used the distributional approximation $(X_{\tau_1}, \cdots, X_{\tau_{M-1}}) \sim N(\mathbf{x}^*, \Sigma^*)$ where

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \ln p(X_{\tau_1}, \cdots, X_{\tau_{M-1}} | X_{\tau_0}, X_{\tau_M}) \tag{51}$$

$$\Sigma^2 = -\left[\frac{\partial^2 \ln p(X^*_{\tau_1}, \cdots, X^*_{\tau_{M-1}} | X_{\tau_0}, X_{\tau_M})}{\partial \mathbf{x}' \partial \mathbf{x}}\right]^{-1}, \tag{52}$$

where $\mathbf{x} = (X_{\tau_1}, \cdots, X_{\tau_{M-1}})'$.

Durham and Gallant (2002) compared the performance of these three importance functions relative to Pedersen (1995) and found that all these methods deliver substantial improvements.

3.4 Other approaches

3.4.1 Numerical ML

While the transition density may not have a closed-form expression for a continuous time model, it must satisfy the Fokker-Planck-Komogorov (also known as "forward") equation. That is,

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial y^2}. \tag{53}$$

where $p(y, t|x, s)$ is the transition density. Solving the partial differential equation numerically at $y = X_{ih}, x = X_{(i-1)h}$ yields the transition density. This is approach proposed by Lo (1988). Similarly, one can numerically solve the "backward" equation

$$\frac{\partial p}{\partial s} = -\frac{1}{2}\frac{\partial^2 p}{\partial x^2}. \tag{54}$$

Obviously, solving these two partial differential equations numerically can be computationally demanding. Consequently, this approach has been little used in practical work.

3.4.2 An exact Gaussian method based on time changes

Yu and Phillips (2001) developed an exact Gaussian method to estimate continuous time models with a linear drift function of the following form:

$$dX(t) = \kappa(\mu - X(t))dt + \sigma(X(t);\theta)dB(t), \tag{55}$$

The approach is based on the idea that any continuous time martingale can be written as a Brownian motion after a suitable time change. That is, when we adjust from chronological time in a local martingale M_t to time based on the evolution of the quadratic variation process $[M]_t$ of M, we have the time change given by $T_t = \inf\{s|[M]_s > t\}$ and the process transforms to a Brownian motion (called DDS Brownian motion) so that $M_t = W_{[M]_t}$, where W is standard Brownian motion.

To see how this approach can be used to estimate equation (55), first write (55) as

$$X(t+\delta) = e^{-\kappa h}X(t) + \mu\left(1 - e^{-\kappa h}\right) + \int_0^{\delta} \sigma e^{-\kappa(\delta-\tau)}\sigma(t+\tau)dB(\tau), \forall\delta > 0. \tag{56}$$

Define $M(\delta) = \sigma\int_0^{\delta} e^{-\kappa(\delta-\tau)}\sigma(t+\tau)dB(\tau)$, which is a continuous martingale with quadratic variation process

$$[M]_{\delta} = \sigma^2 \int_0^{\delta} e^{-2\kappa(\delta-\tau)}\sigma^2(t+\tau)d\tau. \tag{57}$$

To construct a DDS Brownian motion to represent $M(\delta)$, one can construct a sequence of positive numbers $\{\delta_j\}$ which deliver the required time changes. For any fixed constant $a > 0$, let

$$\delta_{j+1} = \inf\{s|[M_j]_s \geq a\} = \inf\{s|\sigma^2 \int_0^{s} e^{-2\kappa(s-\tau)}\sigma^2(t_j+\tau)d\tau \geq a\}. \tag{58}$$

Next, construct a sequence of time points $\{t_j\}$ using the iterations $t_{j+1} = t_j + \delta_{j+1}$ with t_1 assumed to be 0. Evaluating equation (56) at $\{t_j\}$, we have

$$X_{t_{j+1}} = \mu\left(1 - e^{-\kappa\delta_{j+1}}\right) + e^{-\kappa\delta_{j+1}}X_{t_j} + M(\delta_{j+1}). \tag{59}$$

where $M(\delta_{j+1}) = W_{[M]_{\delta_{j+1}}} = W_a \equiv N(0,a)$ is the DDS Brownian motion. Hence, equation (59) is an exact discrete model with Gaussian disturbances and can be estimated directly by ML conditional on the sequence of time

changes. Of course, since the new sequence of time points $\{t_j\}$ is path dependent, this approach does not deliver the true likelihood. Also, since a continuous record of observations is not available, the time points $\{t_j\}$ must be approximated.

4 Approximate ML Methods Based on the Continuous Record Likelihood and Realized Volatility

While (1) is formulated in continuous time, the sample data are always collected at discrete points in time or over discrete intervals in the case of flow data. One may argue that for highly liquid financial assets, the sampled data are so frequently observed as to be nearly continuously available. This is especially true for some tick-by-tick data. Unfortunately, at the highest frequencies, continuous time models such as that given by (1) are often bad descriptions of reality. One reason for the discrepancy is the presence of market microstructure noise, due to trading frictions, bid-ask bounces, recording errors and other anomalies. As a result of these noise effects, the exact ML method based on the continuous record likelihood that was reviewed in Section 2.2 is not applicable.

An alternative approach that is available in such situations was developed in Phillips and Yu (2007) and involves a two-step procedure to estimate the underlying continuous time model that makes use of the empirical quadratic variation process. To explain the method, suppose the model has the form

$$dX(t) = \mu(X(t); \theta_1)dt + \sigma(X(t); \theta_2)dB(t), \tag{60}$$

Note that in this specification the vector of parameters θ_2 in the diffusion function is separated from the parameter vector, θ_1, that appears in the drift function. The reason for this distinction will become clear below.

In the first step, Phillips and Yu (2007) propose to estimate parameters in the diffusion function from the empirical quadratic variation process or so-called realized volatility. The approach is justified by the fact that realized volatility is a natural consistent estimate of quadratic variation and, with certain modifications, can be made consistent even in the presence of microstructure noise effects. Also, realized volatility has convenient distributional characteristics that are determined asymptotically by (functional) central limit theory, as derived by Jacod (1994) and Barndorff-Nielsen and Shephard (2002).

To proceed, assume that X_t is observed at the following times

$$t = \underbrace{h, 2h, \cdots, M_h h(= \frac{T}{K})}, \underbrace{(M_h + 1)h, \cdots, 2M_h h(= \frac{2T}{K})}, \cdots, n_h h(= T), \tag{61}$$

where $n_h = KM_h$ with K a fixed and positive integer, T is the time span of the data, h is the sampling frequency, and $M_h = O(n_h)$. Phillips and Yu constructed the non-overlapping K subsamples

$$((k-1)M_h + 1)h, \cdots, kM_h h, \quad \text{where } k = 1, \cdots, K, \tag{62}$$

so that each sub-sample has M_h observations over the interval $((k-1)\frac{T}{K}, k\frac{T}{K}]$. For example, if ten years of weekly observed data are available and we split the data into ten blocks, then $T = 10$, $h = 1/52$, $M_h = 52$, $K = 10$. The total number of observations is 520 and the number of observations contained in each block is 52.

As $h \to 0$, $n = \frac{T}{h} \to \infty$ and $M_h \to \infty$,

$$\sum_{i=2}^{M_h} (X_{(k-1)M_h+ih} - X_{(k-1)M_h+(i-1)h})^2 \xrightarrow{p} [X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}}, \tag{63}$$

and

$$\frac{\ln(\sum_{i=2}^{M_h} (X_{(k-1)M_h+ih} - X_{(k-1)M_h+(i-1)h})^2 - \ln([X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}}) + \frac{1}{2}s_k^2}{s_k} \xrightarrow{d} N(0,1), \tag{64}$$

where

$$s_k = \min\left\{ \sqrt{\frac{r_k^2}{(\sum_{i=2}^{M_h} (X_{(k-1)M_h+ih} - X_{(k-1)M_h+(i-1)h})^2)^2}}, \sqrt{\frac{2}{M_h}} \right\},$$

$$r_k = \sqrt{\frac{2}{3} \sum_{i=2}^{M_h} (X_{(k-1)M_h+ih} - X_{(k-1)M_h+(i-1)h})^4},$$

for $k = 1, \cdots, K$, and $[X]_T$ is the quadratic variation of X which can be consistently estimated by the empirical counterpart $[X_h]_T$ defined as

$$[X_h]_T = \sum_{i=2}^{n_h} (X_{ih} - X_{(i-1)h})^2. \tag{65}$$

The limit (63) follows by virtue of the definition of quadratic variation. The central limit theorem (CLT) (64) is based on the asymptotic theory of Barndorff-Nielsen and Shephard (2005), which involves a finite sample correction (65) on some important earlier limit theory contributions made by Jacod (1994) and Barndorff-Nielsen and Shephard (2002).

Based on the CLT (64), θ_2 can be estimated in the first stage by running a (nonlinear) least squares regression of the standardized realized volatility

$$\frac{\ln\left(\sum_{i=2}^{M_h}(X_{(k-1)M_h+ih}-X_{(k-1)M_h+(i-1)h})^2\right)+\frac{1}{2}s_k^2}{s_k} \tag{66}$$

on the standardized diffusion function

$$\begin{aligned}\frac{\ln\left([X]_{k\frac{T}{K}}-[X]_{(k-1)\frac{T}{K}}\right)}{s_k} &= \frac{\ln\left(\int_{(k-1)\frac{T}{K}}^{k\frac{T}{K}}\sigma^2\left(X_t;\theta_2\right)dt\right)-\frac{1}{2}s_k^2}{s_k} \\ &\simeq \frac{\ln\left(\sum_{i=2}^{M}\sigma^2\left(X_{(k-1)M_h+(i-1)h};\theta_2\right)h\right)-\frac{1}{2}s_k^2}{s_k}\end{aligned} \tag{67}$$

for $k=1,\cdots,K$. This produces a consistent estimate $\widehat{\theta}_2$ of θ_2. In the second stage, the approximate continuous record or infill log-likelihood function (AIF) is maximized with respect to θ_1

$$\ell_{AIF}(\theta_1)=\sum_{i=2}^{n}\frac{\mu(X_{(i-1)h};\theta_1)}{\sigma^2(X_{(i-1)h};\widehat{\theta}_2)}(X_{ih}-X_{(i-1)h})-\frac{h}{2}\sum_{i=2}^{n}\frac{\mu^2(X_{(i-1)h};\theta_1)}{\sigma^2(X_{(i-1)h};\widehat{\theta}_2)}. \tag{68}$$

The procedure is discussed more fully in Phillips and Yu (2007).

To illustrate the two-stage method, we consider the following specific models.

1. **Vasicek model (6):** Since there is only one parameter in the diffusion function, one could choose $M_h=1$. As a result, the first stage estimation gives the following estimator for σ,

$$\widehat{\sigma}=\sqrt{\frac{[X_h]_T}{T}}, \tag{69}$$

and the approximate infill log-likelihood function is given by

$$\ell_{AIF}(\kappa,\mu)=\sum_{i=2}^{n}\kappa(\mu-X_{(i-1)h})(X_{ih}-X_{(i-1)h})-\frac{h}{2}\sum_{i=2}^{n}\kappa^2(\mu-X_{(i-1)h})^2. \tag{70}$$

2. **Square root model (9):** With $M_h=1$, the first stage estimation gives the following estimator for σ.

$$\widehat{\sigma}=\sqrt{\frac{[X_h]_T}{h\sum_{i=1}^{n_h}X_{(i-1)h}}}. \tag{71}$$

The approximate infill log-likelihood function is given by

$$\ell_{AIF}(\kappa,\mu)=\sum_{i=2}^{n}\frac{\kappa(\mu-X_{(i-1)h})}{\widehat{\sigma}^2X_{(i-1)h}}(X_{ih}-X_{(i-1)h})-\frac{h}{2}\sum_{i=2}^{n}\frac{\kappa^2(\mu-X_{(i-1)h})^2}{\widehat{\sigma}^2X_{(i-1)h}}. \tag{72}$$

5 Monte Carlo Simulations

This section reports the results of a Monte Carlo experiment designed to compare the performance of the various ML estimation methods reviewed in the previous sections. In the experiment, the true generating process is assumed to be the CIR model of short term interest rates of the form

$$dX(t) = \kappa(\mu - X(t))dt + \sigma\sqrt{X(t)}\, dB(t), \tag{73}$$

where $\kappa = 0.1, \mu = 0.1, \sigma = 0.1$. Replications involving 1000 samples, each with 120 monthly observations (ie $h = 1/12$), are simulated from the true model. The parameter settings are realistic to those in many financial applications and the sample period covers 10 years.

It is well-known that κ is difficult to estimate with accuracy whereas the other two parameters, especially σ, are much easier to estimate (Phillips and Yu (2005a, b)) and extensive results are already in the literature. Consequently, we only report estimates of κ in the present Monte Carlo study. In total, we employ six estimation methods, namely, exact ML, the Euler scheme, the Milstein scheme, the Nowman method, the infill method, and the Hermite expansion (with $K = 1$).

Table 1 reports the means, standard errors, and root mean square errors (RMSEs) for all these cases. The exact ML estimator is calculated for comparison purposes. Since the other estimators are designed to approach to the exact ML estimator, we also report the means and the standard errors of the differences between the exact ML estimator and the alternative estimators.

Table 1

Exact and Approximate ML Estimation and Bias Reduced Estimation of κ									
True Value $\kappa = 0.1$									
Method	Exact	Euler	Milstein	Nowman	In-fill	Hermite	Jackk (m=2)	Jackk (m=3)	Ind Inf
Mean	.2403	.2419	.2444	.2386	.2419	.2400	.1465	.1845	.1026
Std err	.2777	.2867	.2867	.2771	.2867	.2762	.3718	.3023	.2593
RMSE	.3112	.3199	.3210	.3098	.3199	.3096	.3747	.3139	.2594
Mean of diff	NA	.0016	.0041	-.0017	.0016	-.0003	NA	NA	NA
Std err of diff	NA	.0500	.0453	.0162	.0500	.0043	NA	NA	NA

Note: A square-root model with $\kappa = 0.1, \mu = 0.1, \sigma = 0.1$ is used to simulate 120 monthly observations for each of the 1,000 replications. Various methods are used to estimate κ.

Several conclusions can be drawn from the table (Note the true value of $\kappa = 0.1$). First, the ML estimator of κ is upward biased by more than 140%,

consistent with earlier results reported in Phillips and Yu (2005a, b). This result is also consistent with what is known about dynamic bias in local-to-unity discrete time autoregressive models. Second, all the approximation-based ML methods perform very similarly to the exact ML method, and hence, all inherit substantial estimation bias from the exact ML method that these methods seek to imitate. Indeed, compared to the estimation bias in exact ML, the bias that is induced purely by the approximations is almost negligible. Third, relative to the Euler scheme, the Milstein scheme fails to offer any improvements in terms of both mean and variation while Nowman's method offers slight improvements in terms of variation and root mean squared error (RMSE). In terms of the quality of approximating the exact ML, the method based on the Hermite expansions is a clear winner when K is as small as 1. Further improvements can be achieved by increasing the value of K, although such improvements do not help to remove the finite sample bias of the ML procedure.

6 Estimation Bias Reduction Techniques

It has frequently been argued in the continuous time finance literature that ML should be the preferred choice of estimation method. The statistical justification for this choice is the generality of the ML approach and its good asymptotic properties of consistency and efficiency. Moreover, since sample sizes in financial data applications are typically large[3], it is often expected that these good asymptotic properties will be realized in finite samples. However, for many financial time series, the asymptotic distribution of the ML estimator often turns out to be a poor approximation to the finite sample distribution, which may be badly biased even when the sample size is large. This is especially the case in the commonly occurring situation of drift parameter estimation in models where the process is nearly a martingale. From the practical viewpoint, this is an important shortcoming of the ML method. The problem of estimation bias turns out to be of even greater importance in the practical use of econometric estimates in asset and option pricing, where there is nonlinear dependence of the pricing functional on the parameter estimates, as shown in Phillips and Yu (2005a). This nonlinearity seems to exacerbate bias and makes good bias correction more subtle.

In the following sections we describe two different approaches to bias correction. The first of these is a simple procedure based on Quenouille's (1956) jackknife. To improve the finite sample properties of the ML estimator in continuous time estimation and in option pricing applications, Phillips and Yu (2005a) proposed a general and computationally inexpensive method of bias reduction based on this approach. The second approach is simulation-

[3] Time series samples of weekly data often exceed 500 and sample sizes are very much larger for daily and intradaily data.

based and involves the indirect inference estimation idea of Gourieroux et al (1993). Monfort (1996) proposed this method of bias corrected estimation in the context of nonlinear diffusion estimation.

In the context of OU process with a known long-run mean, Yu (2007) derived analytical expressions to approximate the bias of ML estimator of the mean reversion parameter and argued that a nonlinear term in the bias formula is particularly important when the mean reversion parameter is close to zero.

6.1 Jackknife estimation

Quenouille (1956) proposed the jackknife as a solution to finite sample bias problems in parametric estimation contexts such as discrete time autoregressions. The method involves the systematic use of subsample estimates. To fix ideas, let N be the number of observations in the whole sample and decompose the sample into m consecutive subsamples each with ℓ observations, so that $N = m \times \ell$. The jackknife estimator of a certain parameter, θ, then utilizes the subsample estimates of θ to assist in the bias reduction process giving the jackknife estimator

$$\widehat{\theta}_{jack} = \frac{m}{m-1}\widehat{\theta}_N - \frac{\sum_{i=1}^{m}\widehat{\theta}_{li}}{m^2 - m}, \tag{74}$$

where $\widehat{\theta}_N$ and $\widehat{\theta}_{li}$ are the estimates of θ obtained by application of a given method like the exact ML or approximate ML to the whole sample and the i'th sub-sample, respectively. Under quite general conditions which ensure that the bias of the estimates $(\widehat{\theta}_N, \widehat{\theta}_{li})$ can be expanded asymptotically in a series of increasing powers of N^{-1}, it can be shown that the bias in the jackknife estimate $\widehat{\theta}_{jack}$ is of order $O(N^{-2})$ rather than $O(N^{-1})$.

The jackknife has several appealing properties. The first advantage is its generality. Unlike other bias reduction methods, such as those based on corrections obtained by estimating higher order terms in an asymptotic expansion of the bias, the jackknife technique does not rely (at least explicitly) on the explicit form of an asymptotic expansion. This means that it is applicable in a broad range of model specifications and that it is unnecessary to develop explicit higher order representations of the bias. A second advantage of the jackknife is that this approach to bias reduction can be used with many different estimation methods, including general methods like the exact ML method whenever it is feasible or approximate ML methods when the exact ML is not feasible. Finally, unlike many other bias correction methods, the jackknife is computationally much cheaper to implement. In fact, the method is not much more time consuming than the initial estimation itself. A draw-

back with jackknife is that it cannot completely remove the bias as it is only designed to decrease the order of magnitude of the bias.

Table 1 reports the results of the jackknife method applied with $m = 2, 3$ based on the same experimental design above. It is clear that the jackknife makes substantial reductions in the bias but this bias reduction comes with an increase in variance. However, a carefully designed jackknife method can reduce the RMSE.

6.2 Indirect inference estimation

The indirect inference (II) procedure, first introduced by Smith (1993), and extended by Gouriéroux, Monfort, and Renault (1993) and Gallant and Tauchen (1996), can be understood as a generalization of the simulated method of moments approach of Duffie and Singleton (1993). It has been found to be a highly useful procedure when the moments and the likelihood function of the true model are difficult to deal with, but the true model is amenable to data simulation. Since many continuous time models are easy to simulate but present difficulties in the analytic derivation of moment functions and likelihood, the indirect inference procedure has some convenient advantages in working with continuous time models in finance. A carefully designed indirect inference estimator can also have good small sample properties, as shown by MacKinnon and Smith (1998) and Gouriéroux, et al. (2000) in the time series context and by Gouriéroux, Phillips and Yu (2007) in the panel context. The method therefore offers some interesting opportunities for bias correction and the improvement of finite sample properties in continuous time estimation.

Without loss of generality, we focus on the OU process. Suppose we need to estimate the parameter κ in the model

$$dX(t) = \kappa(\mu - X(t))dt + \sigma\, dB(t). \tag{75}$$

from observations $\mathbf{x} = \{X_h, \cdots, X_{Nh}\}$. An initial estimator of κ can be obtained, for example, by applying the Euler scheme to $\{X_h, \cdots, X_{Nh}\}$ (call it $\widehat{\kappa}_N$). Such an estimator is inconsistent (due to the discretization error) and may be seriously biased (due to the poor finite sample property of ML in the low κ or near-unit-root case).

The indirect inference method makes use of simulations to remove the discretization bias. It also makes use of simulations to calibrate the bias function and hence requires neither the explicit form of the bias, nor the bias expansion. This advantage seems important when the computation of the bias expression is analytically involved, and it becomes vital when the bias and the first term of the bias asymptotic expansions are too difficult to compute explicitly.

The idea of indirect inference here is as follows. Given a parameter choice κ, we apply the Euler scheme with a much smaller step size than h (say $\delta = h/10$), which leads to

$$\widetilde{X}^k_{t+\delta} = \kappa(\mu - \widetilde{X}^k_t)h + \widetilde{X}^k_t + \sigma\sqrt{\delta}\epsilon_{t+\delta}, \tag{76}$$

where

$$t = \underbrace{0, \delta, \cdots, h(=10\delta)}, \underbrace{h+\delta, \cdots, 2h(=20\delta)}, 2h+\delta, \cdots, Nh. \tag{77}$$

This sequence may be regarded as a nearly exact simulation from the continuous time OU model for small δ. We then choose every $(h/\delta)^{th}$ observation to form the sequence of $\{\widetilde{X}^k_{ih}\}_{i=1}^N$, which can be regarded as data simulated directly from the OU model with the (observationally relevant) step size h.

Let $\widetilde{\mathbf{x}}^k(\kappa) = \{\widetilde{X}^k_h, \cdots, \widetilde{X}^k_{Nh}\}$ be data simulated from the true model, where $k = 1, \cdots, K$ with K being the number of simulated paths. It should be emphasized that it is important to choose the number of observations in $\widetilde{\mathbf{x}}^k(\kappa)$ to be the same as the number of observations in the observed sequence $\mathbf{x}$ for the purpose of the bias calibration. Another estimator of κ can be obtained by applying the Euler scheme to $\{X^k_h, \cdots, X^k_{Nh}\}$ (call it $\widetilde{\kappa}^k_N$). Such an estimator and hence the expected value of them across simulated paths is naturally dependent on the given parameter choice κ.

The central idea in II estimation is to match the parameter obtained from the actual data with that obtained from the simulated data. In particular, the II estimator of κ is defined as

$$\widehat{\kappa}^{II}_{N,K} = \text{argmin}_\kappa \parallel \widehat{\kappa}_N - \frac{1}{K}\sum_{h=1}^{K} \widetilde{\kappa}^k_N(\kappa) \parallel, \tag{78}$$

where $\parallel \cdot \parallel$ is some finite dimensional distance metric. In the case where K tends to infinity, the II estimator is the solution of the limiting extremum problem

$$\widehat{\kappa}^{II}_N = \text{argmin}_\kappa \parallel \widehat{\kappa}_N - E(\widetilde{\kappa}^k_N(\kappa)) \parallel . \tag{79}$$

This limiting extremum problem involves the so-called binding function

$$b_N(\kappa) = E(\widetilde{\kappa}^k_N(\kappa)), \tag{80}$$

which is a finite sample functional relating the bias to κ. In the case where b_N is invertible, the indirect inference estimator is given by

$$\widehat{\kappa}^{II}_N = b_N^{-1}(\widehat{\kappa}_N). \tag{81}$$

The II estimation procedure essentially builds in a small-sample bias correction to parameter estimation, with the bias (in the base estimate, like ML) being computed directly by simulation.

Indirect inference has several advantages for estimating continuous time models. First, it overcomes the inconsistency problem that is common in many approximate ML methods. Second, the indirect inference technique calibrates the bias function via simulation and hence does not require, just like the jackknife method, an explicit form for the bias function or its expansion. Consequently, the method is applicable in a broad range of model specifications. Thirdly, indirect inference can be used with many different estimation methods, including the exact ML method or approximate ML methods, and in doing so will inherit the good asymptotic properties of these base estimators. For instance, it is well known that the Euler scheme offers an estimator which has very small dispersion relative to many consistent estimators and indirect inference applied to it should preserve its good dispersion characteristic while at the same time achieving substantial bias reductions. Accordingly, we expect indirect inference to perform very well in practice and in simulations on the basis of criteria such as RMSE, which take into account central tendency and variation. A drawback with indirect inference is that it is a simulation-based method and can be computationally expensive. However, with the continuing explosive growth in computing power, such a drawback is obviously of less concern

Indirect inference is closely related to median unbiased estimation (MUE) originally proposed by Andrews (1993) in the context of AR models and subsequently applied by Phillips and Yu (2005a) to reduce bias in the mean reversion estimation in the CIR model. While indirect inference uses expectation as the binding function, MUE uses the median as the binding function. Both methods are simulation-based.

Table 1 reports the results of the indirect inference method with $K = 1000$ based on the same experiment discussed earlier. Clearly, indirect inference is very successful in removing bias and the bias reduction is achieved without increasing the variance. As a result, the RMSE is greatly reduced.

7 Multivariate Continuous Time Models

Multivariate systems of stochastic differential equations may be treated in essentially the same manner as univariate models such as (1) and methods such as Euler-approximation-based ML methods and transition density-approximation-based ML methods continue to be applicable. The literature on such extensions is smaller, however, and there are more and more financial data applications of multivariate systems at present; see, for example, Ghysels et al (1996) and Shephard (2005) for reviews of the stochastic volatility literature and Dai and Singleton (2002) for a review of the term structure literature.

One field where the literature on multivariate continuous time econometrics is well developed is macroeconomic modeling of aggregative behavior.

These models have been found to provide a convenient mechanism for embodying economic ideas of cyclical growth, market disequilibrium and dynamic adjustment mechanisms. The models are often constructed so that they are stochastic analogues (in terms of systems of stochastic differential equations) of the differential equations that are used to develop the models in economic theory. The Bergstrom (1966) approximation, discussed in Section 3.1 above, was developed specifically to deal with such multiple equation systems of stochastic equations. Also, the exact discrete time model corresponding to a system of linear diffusions, extending the Vasicek model in Section 2.1, was developed in Phillips (1972, 1974) as the basis for consistent and efficient estimation of structural systems of linear diffusion equations using nonlinear systems estimation and Gaussian ML estimation.

One notable characteristic of such continuous time systems of equations is that there are many across-equation parameter restrictions. These restrictions are typically induced by the manner in which the underlying economic theory (for example, the theory of production involving a parametric production function) affects the formulation of other equations in the model, so that the parameters of one relation (the production relation) become manifest elsewhere in the model (such as wage and price determination, because of the effect of labor productivity on wages). The presence of these across-equation restrictions indicates that there are great advantages to the use of systems procedures, including ML estimation, in the statistical treatment of systems of stochastic differential equations.

While many of the statistical issues already addressed in the treatment of univariate diffusions apply in systems of equations, some new issues do arise. A primary complication is that of aliasing, which in systems of equations leads to an identification problem when a continuous system in estimated by a sequence of discrete observations at sampling interval h. The manifestation of this problem is evident in a system of linear diffusions for an $n-$ vector process $X(t)$ of the form

$$dX(t) = A(\theta_2) X(t) dt + \Sigma(\theta_2) dW(t), \tag{82}$$

where $A = A(\theta)$ is an $n \times n$ coefficient matrix whose elements are dependent on the parameter vector θ_1, $\Sigma = \Sigma(\theta_2)$ is a matrix of diffusion coefficients dependent on the parameter vector θ_2, and $W(t)$ is $n-$ vector standard Brownian motion. The exact discrete model corresponding to this system has the form

$$X_{ih} = e^{hA(\theta_2)} X_{ih} + N\left(0, \int_0^h e^{sA(\theta_2)} \Sigma(\theta_2) e^{sA(\theta_2)'} ds\right), \tag{83}$$

and the coefficient matrix in this discrete time model involves the matrix exponential function $e^{hA(\theta_2)}$. However, there are in general, an infinite number of solutions (A) to the matrix exponential equation

$$e^{hA} = B^0 \tag{84}$$

where $B^0 = e^{hA^0} = e^{hA\left(\theta_2^0\right)}$ and θ_2^0 is the true value of θ_2. In fact, the solutions of the matrix equation (84) all have the form

$$A = A^0 + TQT^{-1}, \tag{85}$$

where T is a matrix that diagonalizes A^0 (so that $T^{-1}AT = \text{diag}(\lambda_1, ..., \lambda_n)$, assuming that A^0 has distinct characteristics roots $\{\lambda_i : i = 1, ..., n\}$), Q is a matrix of the form

$$Q = \frac{2\pi i}{h} \begin{bmatrix} 0 & 0 & 0 \\ 0 & P & 0 \\ 0 & 0 & -P \end{bmatrix}, \tag{86}$$

and P is a diagonal matrix with integers on the diagonal. The multiple solutions of (84) effectively correspond to aliases of A^0.

Fortunately, in this simple system the aliasing problem is not consequential because there are enough restrictions on the form of the system to ensure identifiability. The problem was originally considered in Phillips (1973). In particular, the coefficient matrix $A = A(\theta)$ is real and is further restricted by its dependence on the parameter vector θ. Also, the covariance matrix of the error process $\int_0^h e^{sA(\theta_2)} \Sigma(\theta_2) e^{sA(\theta_2)'} ds$ in the discrete system is real and necessarily positive semi-definite. These restrictions suffice to ensure the identifiability of A^0 in (84), removing the aliasing problem. Discussion and resolution of these issues is given in Phillips (1973) and Hansen and Sargent (1984). Of course, further restrictions may be needed to ensure that θ_1 and θ_2 are identified in $A\left(\theta_1^0\right)$ and $\Sigma\left(\theta_2^0\right)$.

A second complication that arises in the statistical treatment of systems of stochastic differential equations is that higher order systems involve exact discrete systems of the vector autoregressive and moving average type, which have more complicated likelihood functions. A third complication is that the discrete data often involves both stock and flow variables, so that some variables are instantaneously observed (like interest rates) while other variables (like consumption expenditure) are observed as flows (or integrals) over the sampling interval. Derivation of the exact discrete model and the likelihood function in such cases presents further difficulties - see Phillips (1978) and Bergstrom (1984) - and involves complicated submatrix formulations of matrix exponential series. Most of these computational difficulties have now been resolved and Gaussian ML methods have been regularly used in applied research with these continuous time macroeconometric systems. Bergstrom (1996) provides a survey of the subject area and much of the empirical work. A more recent discussion is contained in Bergstrom and Nowman (2006).

8 Conclusions

Research on ML estimation of continuous time systems has been ongoing in the econometric and statistical literatures for more than three decades. But the subject has received its greatest attention in the last decade, as researchers in empirical finance have sought to use these models in practical applications of importance in the financial industry. Among the more significant of these applications have been the analysis of the term structure of interest rates and the pricing of options and other financial derivatives which depend on parameters that occur in the dynamic equations of motion of variables that are most relevant for financial asset prices, such as interest rates. The equations of motion of such variables are typically formulated in terms of stochastic differential equations and so the econometric estimation of such equations has become of critical importance in these applications. We can expect the need for these methods and for improvements in the statistical machinery that is available to practitioners to grow further as the financial industry continues to expand and data sets become richer. The field is therefore of growing importance for both theorists and practitioners.

References

Ahn, D. and Gao, B. (1999): A parametric nonlinear model of term structure dynamics. *Review of Financial Studies* **12**, 721–762.

Aït-Sahalia, Y. (1999): Transition Densities for Interest Rate and Other Nonlinear Diffusions, *Journal of Finance* **54**, 1361–1395.

Aït-Sahalia, Y. (2002): Maximum likelihood estimation of discretely sampled diffusion: A closed-form approximation approach. *Econometrica* **70**, 223–262.

Aït-Sahalia, Y. (2007): Closed-Form Likelihood Expansions for Multivariate Diffusions. *Annals of Statistics* forthcoming.

Aït-Sahalia, Y. and Kimmel, R. (2005): Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions. *Working Paper, Department of Economics, Princeton University.*

Aït-Sahalia, Y. and Kimmel, R. (2007): Maximum Likelihood Estimation of Stochastic Volatility Models. *Journal of Financial Economics* **83**, 413–452.

Aït-Sahalia, Y. and Yu, J. (2006): Saddlepoint approximation for continuous-time Markov Processes. *Journal of Econometrics* **134**, 507–551.

Andrews, D.W.K. (1993): Exactly Median-unbiased Estimation of First Order Autoregressive/unit Root Models. *Econometrica* **61**, 139–166.

Bakshi, G. and Ju, N. (2005): A Refinement to Aït-Sahalia's, 2002 "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach". *Journal of Business* **78**, 2037–2052.

Barndorff-Nielsen, O. and Shephard, N. (2002): Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* **64**, 253–280.

Barndorff-Nielsen, O. and Shephard, N. (2005): How accurate is the asymptotic approximation to the distribution of realized volatility?. *In: Andrews, D.W.K., Powell, J., Ruud, P. and Stock, J. (Eds.): Identification and Inference for Econometric Models.* Cambridge University Press.

Bergstrom, A.R. (1966): Nonrecursive models as discrete approximation to systems of stochastic differential equations, *Econometrica*, 34, 173-82.

Bergstrom, A.R. (1984): Continuous time stochastic models and issues of aggregation over time. *In: Griliches, Z. and Intriligator, M.D. (Eds.): Handbook of Econometrics* **II**. Elsevier Science, Amsterdam.

Bergstrom, A.R. (1996): Survey of continuous time econometrics *In: Barnett, W.A., Gandolfo, G. and Hillinger, C. (Eds.): Dynamic Disequilibrium Modeling*, 3–25. Cambridge University Press.

Bergstrom, A.R. and Nowman, B. (2006): *A Continuous Time Econometric Model of the United Kingdom with Stochastic Trends*. Cambridge University Press.

Billingsley, P. (1961): *Statistical Inference for Markov Processes*. University of Chicago Press.

Black, F. and Scholes, M. (1973): The pricing of options and corporate liabilities. *Journal of Political Economics* **81**, 637–654.

Butler, R. (2007): *Saddlepoint Approximations with Applications*. Cambridge University Press.

Chan, N.H. and Wei, C.Z. (1988): Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16**, 367–401.

Cox, J., Ingersoll, J. and Ross, S. (1985): A theory of the term structure of interest rates. *Econometrica* **53**, 385–407.

Dai, Q. and Singleton, K. (2003): Term Structure Dynamics in Theory and Reality. *Review of Financial Studies* **16**, 631–678.

Daniels, H.E. (1954): Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* **25**, 631–650.

Durham, G. and Gallant, A.R. (2002): Numerical Techniques for Maximum Likelihood Estimation of Continuous-time Diffusion Processes. *Journal of Business and Economic Statistics* **20**, 297–316.

Duffie, D. and Singleton, K.J. (1993): Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* **61**, 929–952.

Elerian, O. (1998): A Note on the Existence of a Closed-form Conditional Transition Density for the Milstein Scheme. *Economics discussion paper 1998-W18, Nuffield College, Oxford.*

Elerian, O., Chib, S. and Shephard, N. (2001): Likelihood inference for discretely observed non-linear diffusions. *Econometrica* **69**, 959–993.

Feller, W. (1951): Two Singular Diffusion Problems. *Annals of Mathematics* **54**, 173–182.

Field, C. and Ronchetti, E. (1990): *Small Sample Asymptotics*. IMS Lecture Notes **13**. Hayward, California.

Gallant, A.R. and Tauchen, G. (1996): Which moments to match? *Econometric Theory* **12**, 657–681.

Ghysels, E., Harvey, A.C. and Renault, E. (1996): Stochastic volatility. *In: Rao, C.R. and Maddala, G.S (Eds.): Statistical Models in Finance*, 119–191. North-Holland, Amsterdam.

Gouriéroux, C., Monfort, A. and Renault, E. (1993): Indirect Inference. *Journal of Applied Econometrics* **8**, 85–118.

Gouriéroux, C., Phillips, P.C.B. and Yu, J. (2007): Indirect inference for dynamic panel models. *Journal of Econometrics* forthcoming.

Gouriéroux, C., Renault, E. and Touzi, N. (2000): Calibration by simulation for small sample bias correction. *In: Mariano, R.S., Schuermann, T. and Weeks, M. (Eds.): Simulation-Based Inference in Econometrics: Methods and Applications*, 328–358. Cambridge University Press.

Hall, P. and Heyde, C.C. (1980): *Martingale Limit Theory and Its Application*. Academic Press.

Hansen, L.P. and Sargent, T.J. (1983): The dimensionality of the aliasing problem in models with rational spectral densities. *Econometrica* **51**, 377–388.

Holly, A. and Phillips, P.C.B. (1979): An saddlepoint approximation to the distribution to the k-class estimator in a simultaneous system. *Econometrica* **47**, 1527–1548.

Houthakker, H.S. and Taylor, L.D. (1966): *Consumer demand in the United States 1929-1970, Analysis and Projections*. Cambridge: Harvard University Press.

Jacod, J. (1994): Limit of random measures associated with the increments of a Brownian semiartingal. *Working paper, Laboratoire de Probabilities, Universite Pierre et Marie Curie, Paris*.

Karatzas, I. and Shreve, S.E. (1991): *Brownian Motion and Stochastic Calculus*. Springer, New York.

Kessler, M. (1997): Estimation of an ergordic diffusion from discrete observations. *Scandinavian Journal of Statistics* **24**, 211–229.

Kloeden, P.E. and Platen, E. (1999): *Numerical Solution of Stochastic Differential Equations*. Springer, New York.

Lánska, V. (1979): Minimum contrast estimation in diffusion processes. *Journal of Applied Probability* **16**, 65–75.

Liptser, R.S. and Shiryaev, A.N. (2000): *Statistics of Random Processes*. Springer, New York.

Lo, A.W. (1988): Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data. *Econometric Theory* **4**, 231–247.

MacKinnon, J.G. and Smith, A.A. (1998): Approximate bias correction in econometrics. *Journal of Econom*terics **85**, 205–230.

McCullagh, P. (1987): *Tensor Methods in Statistics*. London: Chapman and Hall.

Merton, R.C. (1980): On Estimating the Expected Return on the Market: An Exploratory Investigation. *Journal of Financial Economics* **8**, 323–361.

Merton, R.C. (1990): *Continuous-time Finance*. Blackwell, Massachusetts.

Milstein, G.N. (1978): A Method of Second-Order Accuracy Integration of Stochastic Differential Equations. *Theory of Probability and its Applications* **23**, 396–401.

Monfort, A. (1996): A reappraisal of misspecified econometric models. *Econometric Theory* **12**, 597–619.

Nowman, K.B. (1997): Gaussian Estimation of Single-factor Continuous Time Models of the Term Structure of Interest Rates. *Journal of Finance* **52**, 1695–1703.

Pedersen, A. (1995): A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observation. *Scandinavian Journal of Statistics* **22**, 55–71.

Phillips, P.C.B. (1972): The Structural Estimation of a Stochastic Differential Equation System. *Econometrica* **40**, 1021-1041.

Phillips, P.C.B. (1973): The problem of identification in finite parameter continuous time models. *Journal of Econometrics* **1**, 351–362.

Phillips, P.C.B. (1974): The Estimation of Some Continuous Time Models. *Econometrica* **42**, 803–823.

Phillips, P.C.B. (1978): Edgeworth and saddlepoint approximations in a first order noncircular autoregression. *Biometrika* **65**, 91–98.

Phillips, P.C.B. (1980): The exact finite sample density of instrumental variable estimators in an equation with n+1 endogenous variables. *Econometrica* **48**, 861–878.

Phillips, P.C.B. (1984): Marginal densities of instrumental variable estimators in the general single equation case. *Advances in Econometrics* **2**, 1–24.

Phillips, P.C.B. (1987): Time series regression with a unit root. *Econometrica* **55**, 277–301.

Phillips, P.C.B. (1991): Error correction and long run equilibrium in continuous time. *Econometrica* **59**, 967–980.

Phillips, P.C.B. and Magdalinos, T. (2007): Limit theory for moderate deviations from unity. *Journal of Econometrics* **136**, 115–130.

Phillips, P.C.B. and Yu, J. (2005a): Jackknifing bond option prices. *Review of Financial Studies* **18**, 707–742.

Phillips, P.C.B. and Yu, J. (2005b): Comments: A selective overview of nonparametric methods in financial econometrics. *Statistical Science* **20**, 338–343.

Phillips, P.C.B. and Yu, J. (2007): A Two-Stage Realized Volatility Approach to Estimation of Diffusion Processes with Discrete Data. *Journal of Econometrics* forthcoming.

Quenouille, M. H. (1956): Notes on Bias in Estimation. *Biometrika* **43**, 353–360.

Reid, N. (1988): Saddlepoint methods and statistical inference. *Statistical Science* **3**, 213–238.

Sargan, J.D. (1974) Some discrete approximations to continuous time stochastic models. *Journal of the Royal Statistical Society, Series B* **36**, 74–90.

Shoji, I. and Ozaki, T. (1997): Comparative study of estimation methods for continuous time stochastic processes. *Journal of Time Series Analysis* **18**, 485–506.

Shoji, I. and Ozaki, T. (1998): Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications* **16**, 733–752.

Smith, A.A. (1993): Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* **8**, 63–84.

Shephard, N. (2005): *Stochastic Volatility: Selected Readings*. Oxford University Press, Oxford.

Tierney, L. and Kadane, J. (1986): Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–87.

Tse, Y., Zhang, X. and Yu, J. (2004): Estimation of Hyperbolic Diffusion using MCMC Method. *Quantitative Finance* **4**, 158–169.

Vasicek, O. (1977): An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177–186.

Yu, J. (2007): Biases in the Estimation of Mean Reversion Parameter in a Simple Continuous Time Model. *Working Paper, Singapore Management University*.

Yu, J. and Phillips, P.C.B. (2001): A Gaussion Approach for Estimating Continuous Time Models of Short Term Interest Rates. *The Econometrics Journal* **4**, 211–225.

Parametric Inference for Discretely Sampled Stochastic Differential Equations

Michael Sørensen*

Abstract A review is given of parametric estimation methods for discretely sampled multivariate diffusion processes. The main focus is on estimating functions and asymptotic results. Maximum likelihood estimation is briefly considered, but the emphasis is on computationally less demanding martingale estimating functions. Particular attention is given to explicit estimating functions. Results on both fixed frequency and high frequency asymptotics are given. When choosing among the many estimators available, guidance is provided by simple criteria for high frequency efficiency and rate optimality that are presented in the framework of approximate martingale estimating functions.

1 Introduction

In this chapter we consider parametric inference based on observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ from a d-dimensional diffusion process given by

$$dX_t = b(X_t;\theta)dt + \sigma(X_t;\theta)dW_t, \tag{1}$$

where σ is a $d \times d$-matrix and W a d-dimensional standard Wiener process. The drift b and the diffusion matrix σ depend on a parameter θ which varies in a subset Θ of $\mathbb{R}^p$. The main focus is on estimating functions and asymptotic results.

Michael Sørensen
Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark, e-mail: michael@math.ku.dk

* The research was supported by the Danish Center for Accounting and Finance funded by the Danish Social Science Research Council and by the Center for Research in Econometric Analysis of Time Series funded by the Danish National Research Foundation.

DOI: 10.1007/978-3-540-71297-8_23,

The true (data-generating) model is supposed to be the stochastic differential equation (1) with the parameter value θ_0, and the coefficients b and σ are assumed to be sufficiently smooth functions of the state to ensure the existence of a unique weak solution for all θ in Θ. The state space of X is denoted by D. When $d = 1$, the state space is an interval (ℓ, r), where ℓ could possibly be $-\infty$, and r might be ∞. We suppose that the transition distribution has a density $y \mapsto p(\Delta, x, y; \theta)$ with respect to the Lebesgue measure on D, and that $p(\Delta, x, y; \theta) > 0$ for all $y \in D$. The transition density is the conditional density of $X_{t+\Delta}$ given that $X_t = x$. Since the data are equidistant, we will often suppress the argument Δ in the transition density and write $p(x, y; \theta)$.

It is assumed that the diffusion is ergodic, and that its invariant probability measure has density function μ_θ for all $\theta \in \Theta$. The initial value of the diffusion is assumed to be either known, $X_0 = x_0$, or $X_0 \sim \mu_\theta$. In the latter case the diffusion is stationary.

2 Asymptotics: Fixed Frequency

We consider the asymptotic properties of an estimator $\widehat{\theta}_n$ obtained by solving the estimating equation

$$G_n(\widehat{\theta}_n) = 0, \tag{2}$$

where G_n is an estimating function of the form

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta, X_{\Delta i}, X_{\Delta(i-1)}; \theta) \tag{3}$$

for some suitable function $g(\Delta, y, x; \theta)$ with values in $\mathbb{R}^p$. All estimators discussed below can be represented in this way. An estimator, $\widehat{\theta}_n$, which solves (2) with probability approaching one as $n \to \infty$, is called a G_n–estimator. A priori there is no guarantee that a unique solution to (2) exists. In this section, we consider the standard asymptotic scenario, where the time between observations Δ is fixed and the number of observations goes to infinity. In most cases we suppress Δ in the notation and write for example $g(y, x; \theta)$.

We have assumed that the diffusion is ergodic and denote the density function of the invariant probability measure by μ_θ. Let Q_θ denote the probability measure on D^2 with density function $\mu_\theta(x)p(\Delta, x, y; \theta)$. This is the density function of two consecutive observations $(X_{\Delta(i-1)}, X_{\Delta i})$ when the diffusion is stationary, i.e. when $X_0 \sim \mu_\theta$. We impose the following condition on the function g

$$Q_\theta(g_j(\theta)^2) = \int_{D^2} g_j(y, x; \theta)^2 \mu_\theta(x) p(x, y; \theta) dy dx < \infty, \quad j = 1, \ldots, p, \tag{4}$$

for all $\theta \in \Theta$, where g_j denotes the jth coordinate of g. The quantity $Q_\theta(g_j(\theta))$ is defined similarly. Under the assumption of ergodicity and (4), it follows that

$$\frac{1}{n}\sum_{i=1}^{n} g(X_{\Delta i}, X_{\Delta(i-1)}; \theta) \xrightarrow{P_\theta} Q_\theta(g(\theta))^2. \tag{5}$$

When the diffusion, X, is one-dimensional, the following simple conditions ensure *ergodicity*, and an explicit expression exists for the density of the invariant probability measure. The *scale measure* of X has Lebesgue density

$$s(x;\theta) = \exp\left(-2\int_{x^\#}^{x} \frac{b(y;\theta)}{\sigma^2(y;\theta)}dy\right), \quad x \in (\ell, r), \tag{6}$$

where $x^\# \in (\ell, r)$ is arbitrary.

Condition 1 *The following holds for all $\theta \in \Theta$:*

$$\int_{x^\#}^{r} s(x;\theta)dx = \int_{\ell}^{x^\#} s(x;\theta)dx = \infty$$

and

$$\int_{\ell}^{r} [s(x;\theta)\sigma^2(x;\theta)]^{-1}dx = A(\theta) < \infty.$$

Under Condition 1 the process X is ergodic with an invariant probability measure with Lebesgue density

$$\mu_\theta(x) = [A(\theta)s(x;\theta)\sigma^2(x;\theta)]^{-1}, \quad x \in (\ell, r). \tag{7}$$

For details see e.g. Skorokhod (1989).

For the following asymptotic results to hold, we also need to assume that under P_θ the estimating function (3) satisfies a central limit theorem

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_{\Delta i}, X_{\Delta(i-1)}; \theta) \xrightarrow{\mathcal{D}} N(0, V(\theta)) \tag{8}$$

for some $p \times p$-matrix $V(\theta)$. For (8) to hold, it is obviously necessary that $Q_\theta(g(\theta)) = 0$.

Theorem 1 *Assume that $\theta_0 \in \operatorname{int}\Theta$ and that a neighbourhood N of θ_0 in Θ exists, such that:*
(1) The function $g(\theta) : (x, y) \mapsto g(x, y; \theta)$ is integrable with respect to the probability measure Q_{θ_0} for all $\theta \in N$, and

$$Q_{\theta_0}(g(\theta_0)) = 0. \tag{9}$$

[2] $Q_\theta(g(\theta))$ denotes the vector $(Q_\theta(g_j(\theta)))_{j=1,\ldots,p}$.

(2) The function $\theta \mapsto g(x,y;\theta)$ *is continuously differentiable on* N *for all* $(x,y) \in D^2$.
(3) The functions[3] $(x,y) \mapsto \partial_{\theta_j} g_i(x,y;\theta)$, $i,j = 1,\ldots,p$, *are dominated for all* $\theta \in N$ *by a function which is integrable with respect to* Q_{θ_0}.
(4) The $p \times p$ *matrix*[4]

$$W = Q_{\theta_0}\left(\partial_{\theta^T} g(\theta_0)\right) \tag{10}$$

is invertible.

Then a consistent G_n*–estimator* $\widehat{\theta}_n$ *exists, and*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\left(0, W^{-1}VW^{T^{-1}}\right) \tag{11}$$

under P_{θ_0}, *where* $V = V(\theta_0)$. *If, moreover, the function* $g(x,y;\theta)$ *is locally dominated integrable*[5] *with respect to* Q_{θ_0} *and*

$$Q_{\theta_0}(g(\theta)) \neq 0 \quad \textit{for all} \quad \theta \neq \theta_0,$$

then the estimator $\widehat{\theta}_n$ *is unique on any bounded subset of* Θ *containing* θ_0 *with probability approaching one as* $n \to \infty$.

A proof of this theorem can be found in Jacod and Sørensen (2008). Related asymptotic results formulated in the language of the generalized method of moments were given by Hansen (1982).

If an estimating function does not satisfy (9), the obtained estimator is not consistent, but will converge to the solution $\bar{\theta}$ to the equation

$$Q_{\theta_0}(g(\bar{\theta})) = 0. \tag{12}$$

If the estimating function $G_n(\theta)$ is a martingale under P_θ, the asymptotic normality in (8) follows without further conditions from the central limit theorem for martingales, see Hall and Heyde (1980). This result goes back to Billingsley (1961). In the martingale case the matrix $V(\theta)$ is given by

$$V(\theta) = Q_{\theta_0}\left(g(\theta)g(\theta)^T\right), \tag{13}$$

and the asymptotic covariance matrix of the estimator $\widehat{\theta}_n$ can be consistently estimated by means of the matrices W_n and V_n given in the following theorem; see Jacod and Sørensen (2008).

[3] $\partial_{\theta_j} g_i$ denotes the partial derivative $\frac{\partial g_i}{\partial \theta_j}$.

[4] In this chapter T denotes transposition, vectors are column vectors, and $Q_{\theta_0}\left(\partial_{\theta^T} g(\theta_0)\right)$ denotes the matrix $\{Q_{\theta_0}(\partial_{\theta_j} g_i(\theta_0))\}$, where i is the row number and j the column number.

[5] A function $g : D^2 \times \Theta \mapsto \mathbb{R}$ is called locally dominated integrable with respect to Q_{θ_0} if for each $\theta' \in \Theta$ there exists a neighbourhood $U_{\theta'}$ of θ' and a non-negative Q_{θ_0}-integrable function $h_{\theta'} : D^2 \mapsto \mathbb{R}$ such that $|\, g(x,y;\theta)\,| \leq h_{\theta'}(x,y)$ for all $(x,y,\theta) \in D^2 \times U_{\theta'}$.

Theorem 2 *Under the conditions (2) – (4) of Theorem 1,*

$$W_n = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta g(X_{(i-1)\Delta}, X_{i\Delta}; \widehat{\theta}_n) \xrightarrow{P_{\theta_0}} W, \tag{14}$$

and the probability that W_n is invertible approaches one as $n \to \infty$. If, moreover, the functions $(x, y) \mapsto g_i(x, y; \theta)$, $i = 1, \ldots, p$, are dominated for all $\theta \in N$ by a function which is square integrable with respect to Q_{θ_0}, then in the martingale case

$$V_n = \frac{1}{n}\sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}; \widehat{\theta}_n) g(X_{(i-1)\Delta}, X_{i\Delta}; \widehat{\theta}_n)^T \xrightarrow{P_{\theta_0}} V. \tag{15}$$

When the estimating function $G_n(\theta)$ is not a martingale under P_θ, further conditions on the diffusion process must be imposed to ensure the asymptotic normality in (8). If the diffusion is stationary and geometrically α-mixing[6], (8) holds with

$$V(\theta) = Q_{\theta_0}\left(g(\theta)g(\theta)^T\right) + \sum_{k=1}^{\infty} \left[E_{\theta_0}\left(g(X_\Delta, X_0) g(X_{(k+1)\Delta}, X_{k\Delta})^T\right) + E_{\theta_0}\left(g(X_{(k+1)\Delta}, X_{k\Delta}) g(X_\Delta, X_0)^T\right)\right], \tag{16}$$

provided that $V(\theta)$ is strictly positive definite, and that $Q_\theta(g_i(\theta)^{2+\epsilon}) < \infty$ for some $\epsilon > 0$, see e.g. Doukhan (1994). Genon-Catalot et al. (2000) gave the following simple sufficient condition for a one-dimensional diffusion to be geometrically α-mixing.

Condition 2
(i) The function b is continuously differentiable with respect to x and σ is twice continuously differentiable with respect to x, $\sigma(x; \theta) > 0$ for all $x \in (\ell, r)$, and there exists a constant $K_\theta > 0$ such that $|b(x; \theta)| \le K_\theta(1 + |x|)$ and $\sigma^2(x; \theta) \le K_\theta(1 + x^2)$ for all $x \in (\ell, r)$.

(ii) $\sigma(x; \theta)\mu_\theta(x) \to 0$ as $x \downarrow \ell$ and $x \uparrow r$.

(iii) $1/\gamma(x; \theta)$ has a finite limit as $x \downarrow \ell$ and $x \uparrow r$, where $\gamma(x; \theta) = \partial_x \sigma(x; \theta) - 2b(x; \theta)/\sigma(x; \theta)$.

Other conditions for geometric α-mixing were given by Veretennikov (1987), Hansen and Scheinkman (1995), and Kusuoka and Yoshida (2000).

[6] α-mixing with mixing coefficients that tend to zero geometrically fast.

3 Likelihood Inference

The diffusion process X is a Markov process, so the likelihood function based on the observations $X_{t_0}, X_{t_1}, \cdots, X_{t_n}$ $(t_0 = 0)$, conditional on X_0, is

$$L_n(\theta) = \prod_{i=1}^{n} p(t_i - t_{i-1}, X_{t_{i-1}}, X_{t_i}; \theta), \tag{17}$$

where $y \mapsto p(s, x, y; \theta)$ is the transition density. Under weak regularity conditions the maximum likelihood estimator is efficient, i.e. it has the smallest asymptotic variance among all estimators. The transition density is only rarely explicitly known, but several numerical approaches make likelihood inference feasible for diffusion models. Pedersen (1995) proposed a method for obtaining an approximation to the likelihood function by rather extensive simulation. Pedersen's method was very considerably improved by Durham and Gallant (2002), whose method is computationally much more efficient. Poulsen (1999) obtained an approximation to the transition density by numerically solving a partial differential equation, whereas Aït-Sahalia (2002, 2003) proposed to approximate the transition density by means of expansions. A Gaussian approximation to the likelihood function obtained by local linearization of (1) was proposed by Ozaki (1985), while Forman and Sørensen (2008) proposed to use an approximation in terms of eigenfunctions of the generator of the diffusion. Bayesian estimators with the same asymptotic properties as the maximum likelihood estimator can be obtained by Markov chain Monte Carlo methods, see Elerian et al. (2001), Eraker (2001), and Roberts and Stramer (2001). Finally, exact and computationally efficient likelihood-based estimation methods were presented by Beskos et al. (2006). These various approaches to maximum likelihood estimation will not be considered further in this chapter. Some of them are treated in Phillips and Yu (2008). Asymptotic results for the maximum likelihood estimator were established by Dacunha-Castelle and Florens-Zmirou (1986), while asymptotic results when the observations are made at random time points were obtained by Aït-Sahalia and Mykland (2003).

The vector of partial derivatives of the log-likelihood function with respect to the coordinates of θ,

$$U_n(\theta) = \partial_\theta \log L_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta), \tag{18}$$

where $\Delta_i = t_i - t_{i-1}$, is called the *score function* (or score vector). The maximum likelihood estimator usually solves the estimating equation $U_n(\theta) = 0$. The score function is a martingale under P_θ, which is easily seen provided that the following interchange of differentiation and integration is allowed:

$$E_\theta\left(\partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta) \middle| X_{t_1}, \ldots, X_{t_{i-1}}\right)$$

$$= \int_D \frac{\partial_\theta p(\Delta_i, X_{t_{i-1}}, y; \theta)}{p(\Delta_i, X_{t_{i-1}}, y; \theta)} p(\Delta_i, X_{t_{i-1}}, y, \theta) dy$$

$$= \partial_\theta \int_D p(\Delta_i, X_{t_{i-1}}, y; \theta) dy = 0.$$

A simple approximation to the likelihood function is obtained by approximating the transition density by a Gaussian density with the correct first and second conditional moments. For a one-dimensional diffusion we get

$$p(\Delta, x, y; \theta) \approx q(\Delta, x, y; \theta) = \frac{1}{\sqrt{2\pi\phi(\Delta, x; \theta)}} \exp\left[\frac{(y - F(\Delta, x; \theta))^2}{2\phi(\Delta, x; \theta)}\right]$$

where

$$F(\Delta, x; \theta) = \mathrm{E}_\theta(X_\Delta | X_0 = x) = \int_\ell^r y p(\Delta, x, y; \theta) dy. \tag{19}$$

and

$$\phi(\Delta, x; \theta) = \mathrm{Var}_\theta(X_\Delta | X_0 = x) = \int_\ell^r [y - F(\Delta, x; \theta)]^2 p(\Delta, x, y; \theta) dy. \tag{20}$$

In this way we obtain the *quasi-likelihood*

$$L_n(\theta) \approx QL_n(\theta) = \prod_{i=1}^n q(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta),$$

and by differentiation with respect to the parameter vector, we obtain the quasi-score function

$$\partial_\theta \log QL_n(\theta) = \sum_{i=1}^n \left\{ \frac{\partial_\theta F(\Delta_i, X_{t_{i-1}}; \theta)}{\phi(\Delta_i, X_{t_{i-1}}; \theta)} [X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta)] \right. \tag{21}$$

$$\left. + \frac{\partial_\theta \phi(\Delta_i, X_{t_{i-1}}; \theta)}{2\phi(\Delta_i, X_{t_{i-1}}; \theta)^2} \left[(X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta))^2 - \phi(\Delta_i, X_{t_{i-1}}; \theta)\right] \right\},$$

which is clearly a martingale under P_θ. It is a particular case of the quadratic martingale estimating functions considered by Bibby and Sørensen (1995) and Bibby and Sørensen (1996). Maximum quasi-likelihood estimation was considered by Bollerslev and Wooldridge (1992).

4 Martingale Estimating Functions

In this section we present a rather general way of obtaining approximations to the score function by means of martingales of a similar form. Suppose we have a collection of real valued functions $h_j(x, y, ; \theta)$, $j = 1, \ldots, N$ satisfying

$$\int_D h_j(x, y; \theta) p(x, y; \theta) dy = 0 \tag{22}$$

for all $x \in D$ and $\theta \in \Theta$. Each of the functions h_j could be used separately to define an estimating function of the form (3), but a better approximation to the score function, and hence a more efficient estimator, is obtained by combining them in an optimal way. Therefore we consider estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} a(X_{(i-1)\Delta}, \theta) h(X_{(i-1)\Delta}, X_{i\Delta}; \theta), \tag{23}$$

where $h = (h_1, \ldots, h_N)^T$, and the $p \times N$ weight matrix $a(x, \theta)$ is a function of x such that (23) is P_θ-integrable. It follows from (22) that $G_n(\theta)$ is a martingale under P_θ for all $\theta \in \Theta$. An estimating function with this property is called a *martingale estimating function*.

The matrix a determines how much weight is given to each of the h_js in the estimation procedure. This weight matrix can be chosen in an optimal way rather straightforwardly using the theory of optimal estimating functions, see Godambe (1960), Durbin (1960), Godambe and Heyde (1987) and Heyde (1997). The optimal weight matrix a^* gives the estimating function of the form (23) that provides the best possible approximation to the score function (18) in a mean square sense. Moreover, the optimal $g^*(x, y; \theta) = a^*(x; \theta) h(x, y; \theta)$ is obtained from $\partial_\theta \log p(x, y; \theta)$ by projection in a certain space of square integrable functions, see Kessler (1996) and Sørensen (1997).

The choice of the functions h_j, on the other hand, is an art rather than a science. The ability to tailor these functions to a given model or to particular parameters of interest is a considerable strength of the estimating functions methodology. It is, however, also a source of weakness, since it is not always clear how best to choose the h_js. In this and the next section, we shall present ways of choosing these functions that usually work well in practice.

Example 1 The martingale estimating function (21) is of the type (23) with $N = 2$ and

$$\begin{aligned} h_1(x, y; \theta) &= y - F(\Delta, x; \theta), \\ h_2(x, y; \theta) &= (y - F(\Delta, x; \theta))^2 - \phi(\Delta, x, \theta), \end{aligned}$$

where F and ϕ are given by (19) and (20). The weight matrix is

$$\left(\frac{\partial_\theta F(\Delta, x; \theta)}{\phi(\Delta, x; \theta)}, \quad \frac{\partial_\theta \phi(\Delta, x; \theta)}{2\phi^2(\Delta, x; \theta)\Delta} \right), \tag{24}$$

which we shall see is approximately optimal. □

In the econometrics literature, a popular way of using functions like $h_j(x, y, ; \theta)$, $j = 1, \ldots, N$, to estimate the parameter θ is the *generalized method of moments* (GMM) of Hansen (1982). The method is usually implemented as follows, see e.g. Campbell et al. (1997). Consider

$$F_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} h(X_{(i-1)\Delta}, X_{i\Delta}; \theta).$$

Under weak conditions, cf. Theorem 2, a consistent estimator of the asymptotic covariance matrix M of $\sqrt{n} F_n(\theta)$ is

$$M_n = \frac{1}{n} \sum_{i=1}^{n} h(X_{(i-1)\Delta}, X_{i\Delta}; \widetilde{\theta}_n) h(X_{(i-1)\Delta}, X_{i\Delta}; \widetilde{\theta}_n)^T,$$

where $\widetilde{\theta}_n$ is a consistent estimator of θ (for instance obtained by minimizing $F_n(\theta)^T F_n(\theta)$). The GMM-estimator is obtained by minimizing the function

$$H_n(\theta) = F_n(\theta)^T M_n^{-1} F_n(\theta).$$

The corresponding estimating function is obtained by differentiation with respect to θ

$$\partial_\theta H_n(\theta) = D_n(\theta) M_n^{-1} F_n(\theta),$$

where by (5)

$$D_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \partial_\theta h(X_{(i-1)\Delta}, X_{i\Delta}; \theta)^T \xrightarrow{P_{\theta_0}} Q_{\theta_0} \left(\partial_\theta h(\theta)^T \right).$$

Hence the estimating function $\partial_\theta H_n(\theta)$ is asymptotically equivalent to an estimating function of the form (23) with a constant weight matrix

$$a(x, \theta) = Q_{\theta_0} \left(\partial_\theta h(\theta)^T \right) M^{-1},$$

and we see that GMM-estimators are covered by the theory for martingale estimating functions presented in this chapter.

We now return to the problem of finding the optimal estimating function $G_n^*(\theta)$, i.e. of the form (23) with the *optimal weight matrix*. To do so we assume that the functions h_j satisfy the following condition.

convergence

(1) The functions h_j, $j = 1, \ldots N$, are linearly independent.

(2) The functions $y \mapsto h_j(x, y; \theta)$, $j = 1, \ldots N$, are square integrable with respect to $p(x, y; \theta)$ for all $x \in D$ and $\theta \in \Theta$.

(3) $h_j(x, y; \theta)$, $j = 1, \ldots N$, are differentiable with respect to θ.
(4) The functions $y \mapsto \partial_\theta h_j(x, y; \theta)$ are integrable with respect to $p(x, y; \theta)$ for all $x \in D$ and $\theta \in \Theta$.

According to the theory of optimal estimating functions, the optimal choice of the weight matrix a is given by

$$a^*(x; \theta) = B_h(x; \theta)\, V_h(x; \theta)^{-1}, \tag{25}$$

where

$$B_h(x; \theta) = \int_D \partial_\theta h(x, y; \theta)^T p(x, y; \theta) dy \tag{26}$$

and

$$V_h(x; \theta) = \int_D h(x, y; \theta) h(x, y; \theta)^T p(x, y; \theta) dy. \tag{27}$$

The asymptotic variance of an optimal estimator, i.e. a G_n^*–estimator, is simpler than the general expression in (11) because in this case the matrices W and V given by (10) and (13) are equal and given by (29), as can easily be verified. Thus we have the following corollary to Theorem 1:

Corollary 1 *Assume that $g^*(x, y, \theta) = a^*(x; \theta) h(x, y; \theta)$ satisfies the conditions of Theorem 1. Then a sequence $\widehat{\theta}_n$ of G_n^*–estimators has the asymptotic distribution*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\left(0, V^{-1}\right), \tag{28}$$

where

$$V = \mu_{\theta_0}\left(B_h(\theta_0) V_h(\theta_0)^{-1} B_h(\theta_0)^T\right) \tag{29}$$

with B_h and V_h given by (26) *and* (27).

Example 2 Consider the martingale estimating function of form (23) with $N = 2$ and with h_1 and h_2 as in Example 1, where the diffusion is one-dimensional. The optimal weight matrix has columns given by

$$a_1^*(x; \theta) = \frac{\partial_\theta \phi(x; \theta) \eta(x; \theta) - \partial_\theta F(x; \theta) \psi(x; \theta)}{\phi(x; \theta) \psi(x; \theta) - \eta(x; \theta)^2}$$
$$a_2^*(x; \theta) = \frac{\partial_\theta F(x; \theta) \eta(x; \theta) - \partial_\theta \phi(x; \theta) \phi(x; \theta)}{\phi(x; \theta) \psi(x; \theta) - \eta(x; \theta)^2},$$

where

$$\eta(x; \theta) = \mathrm{E}_\theta([X_\Delta - F(x; \theta)]^3 | X_0 = x)$$

and

$$\psi(x; \theta) = \mathrm{E}_\theta([X_\Delta - F(x; \theta)]^4 | X_0 = x) - \phi(x; \theta)^2.$$

We can simplify these expressions by making the Gaussian approximations

$$\eta(t, x; \theta) \approx 0 \quad \text{and} \quad \psi(t, x; \theta) \approx 2\phi(t, x; \theta)^2. \tag{30}$$

If we insert these approximations into the expressions for a_1^* and a_2^*, we obtain the weight functions in (21). When Δ is not large this can be justified, because the transition distribution is not far from Gaussian. □

In the next subsection we shall present a class of martingale estimating functions for which the matrices $B_h(x;\theta)$ and $V_h(x;\theta)$ can be found explicitly, but for most models these matrices must be found by simulation. If a^* is determined by a relatively time consuming numerical method, it might be preferable to use the estimating function

$$G_n^\bullet(\theta) = \sum_{i=1}^n a^*(X_{(i-1)\Delta};\widetilde{\theta}_n)h(X_{(i-1)\Delta}, X_{i\Delta};\theta), \tag{31}$$

where $\widetilde{\theta}_n$ is a weakly $\sqrt{n}$-consistent estimator of θ, for instance obtained by some simple choice of the weight matrix a. In this way a^* needs to be calculated only once per observation point. Under weak regularity conditions, the estimator obtained from $G_n^\bullet(\theta)$ has the same efficiency as the optimal estimator; see e.g. Jacod and Sørensen (2008).

Most martingale estimating functions proposed in the literature are of the form

$$G_n(\theta) = \sum_{i=1}^n a(X_{(i-1)\Delta},\theta)\left[f(X_{i\Delta};\theta) - \pi_\Delta^\theta(f(\theta))(X_{(i-1)\Delta})\right], \tag{32}$$

where $f = (f_1,\ldots,f_N)^T$, and π_Δ^θ denotes the *transition operator*

$$\pi_s^\theta(f)(x) = \int_D f(y)p(s,x,y;\theta)dy = \mathrm{E}_\theta(f(X_s)\,|\,X_0 = x). \tag{33}$$

The polynomial estimating functions given by $f_j(y) = y^j$, $j = 1,\ldots,N$, are an example. For martingale estimating functions of the special form (32), the expression for the optimal weight matrix simplifies to some extent to

$$B_h(x;\theta)_{ij} = \pi_\Delta^\theta(\partial_{\theta_i} f_j(\theta))(x) - \partial_{\theta_i}\pi_\Delta^\theta(f_j(\theta))(x), \tag{34}$$

$i = 1,\ldots p$, $j = 1,\ldots,N$, and

$$V_h(x;\theta)_{ij} = \pi_\Delta^\theta(f_i(\theta)f_j(\theta))(x) - \pi_\Delta^\theta(f_i(\theta))(x)\pi_\Delta^\theta(f_j(\theta))(x), \tag{35}$$

$i,j = 1,\ldots,N$. Often the functions f_j can be chosen such that they do not depend on θ, in which case

$$B_h(x;\theta)_{ij} = -\partial_{\theta_i}\pi_\Delta^\theta(f_j)(x). \tag{36}$$

A useful approximations to the optimal weight matrix can be obtained by applying the formula

$$\pi_s^\theta(f)(x) = \sum_{i=0}^{k} \frac{s^i}{i!} A_\theta^i f(x) + O(s^{k+1}), \tag{37}$$

where A_θ denotes the generator of the diffusion

$$A_\theta f(x) = \sum_{k=1}^{d} b_k(x;\theta)\partial_{x_k} f(x) + \frac{1}{2}\sum_{k,\ell=1}^{d} C_{k\ell}(x;\theta)\partial^2_{x_k x_\ell} f(x), \tag{38}$$

where $C = \sigma\sigma^T$. The formula (37) holds for $2(k+1)$ times continuously differentiable functions under weak conditions which ensure that the remainder term has the correct order, see Kessler (1997). It is often enough to use the approximation $\pi_\Delta^\theta(f_j)(x) \approx f_j(x) + \Delta A_\theta f_j(x)$. When f does not depend on θ this implies that

$$B_h(x;\theta) \approx \Delta\left[\partial_\theta b(x;\theta) f'(x) + \tfrac{1}{2}\partial_\theta \sigma^2(x;\theta) f''(x)\right] \tag{39}$$

and (for $N = 1$)

$$V_h(x;\theta) \approx \Delta\left[A_\theta(f^2)(x) - 2f(x)A_\theta f(x)\right] = \Delta\,\sigma^2(x;\theta) f'(x)^2. \tag{40}$$

Example 3 If we simplify the optimal weight matrix found in Example 2 by (37) and the Gaussian approximation (30), we obtain the approximately optimal quadratic martingale estimating function

$$\begin{aligned} G_n^\circ(\theta) = \sum_{i=1}^{n} \Bigg\{ & \frac{\partial_\theta b(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)}[X_{i\Delta} - F(X_{(i-1)\Delta};\theta)] \\ & + \frac{\partial_\theta \sigma^2(X_{(i-1)\Delta};\theta)}{2\sigma^4(X_{(i-1)\Delta};\theta)\Delta}\left[(X_{i\Delta} - F(X_{(i-1)\Delta};\theta))^2 - \phi(X_{(i-1)\Delta};\theta)\right]\Bigg\}. \end{aligned} \tag{41}$$

For the CIR-model

$$dX_t = -\beta(X_t - \alpha)dt + \tau\sqrt{X_t}dW_t, \tag{42}$$

where $\beta, \tau > 0$, the approximately optimal quadratic martingale estimating function is

$$\begin{pmatrix} \sum_{i=1}^{n} \frac{1}{X_{(i-1)\Delta}} \left[X_{i\Delta} - X_{(i-1)\Delta} e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta}) \right] \\ \sum_{i=1}^{n} \left[X_{i\Delta} - X_{(i-1)\Delta} e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta}) \right] \\ \sum_{i=1}^{n} \frac{1}{X_{(i-1)\Delta}} \left[\left(X_{i\Delta} - X_{(i-1)\Delta} e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta}) \right)^2 \right. \\ \left. \quad - \frac{\tau^2}{\beta} \left\{ \left(\alpha/2 - X_{(i-1)\Delta} \right) e^{-2\beta\Delta} - (\alpha - X_{(i-1)\Delta}) e^{-\beta\Delta} + \alpha/2 \right\} \right] \end{pmatrix}. \tag{43}$$

This is obtained from (41) after multiplication by an invertible non-random matrix to obtain a simpler expression. This does not change the estimator. From this estimating function explicit estimators can easily be obtained. A simulation study and an investigation of the asymptotic variance of the estimators for α and β in Bibby and Sørensen (1995) show that they are not much less efficient than the estimators from the optimal estimating function; see also the simulation study in Overbeck and Rydén (1997). □

When the optimal weight matrix is approximated by means of (37), there is a certain loss of efficiency, which as in the previous example is often quite small; see Bibby and Sørensen (1995) and the section on high frequency asymptotics below. Therefore the relatively simple estimating function (41) is often a good choice in practice.

It is tempting to go on to approximate $\pi_\Delta^\theta(f_j(\theta))(x)$ in (32) by (37) in order to obtain an explicit estimating function, but as we shall see in the next section this is often a dangerous procedure. In general the conditional expectation in π_Δ^θ should therefore be approximated by simulations. Fortunately, Kessler and Paredes (2002) have established that, provided the simulation is done with sufficient accuracy, this does not cause any bias, only a minor loss of efficiency that can be made arbitrarily small. Moreover, as we shall also see in the next section, $\pi_\Delta^\theta(f_j(\theta))(x)$ can be found explicitly for a quite flexible class of diffusions.

5 Explicit Inference

In this section we consider one-dimensional diffusion models for which estimation is particularly easy because an explicit martingale estimating function exists.

Kessler and Sørensen (1999) proposed estimating functions of the form (32) where the functions f_j, $i = 1, \ldots, N$ are *eigenfunctions* for the generator (38), i.e.

$$A_\theta f_j(x;\theta) = -\lambda_j(\theta) f_j(x;\theta),$$

where the real number $\lambda_j(\theta) \geq 0$ is called the *eigenvalue* corresponding to $f_j(x;\theta)$. Under weak regularity conditions, f_j is also an eigenfunction for the transition operator π_t^θ, i.e.

$$\pi_t^\theta(f_j(\theta))(x) = e^{-\lambda_j(\theta)t} f_j(x;\theta). \tag{44}$$

for all $t > 0$. Thus explicit martingales are obtained. Each of the following three conditions imply (44):

(i) $\sigma(x;\theta)$ and $\partial_x f_j(x;\theta)$ are bounded functions of $x \in (\ell, r)$.

(ii) $\int_\ell^r [\partial_x f_j(x;\theta)\sigma(x;\theta)]^2 \mu_\theta(dx) < \infty$.

(iii) b and σ are of linear growth, and $\partial_x f_j$ is of polynomial growth in $x \in (\ell, r)$.

Example 4 The model

$$dX_t = -\beta[X_t - (m + \gamma z)]dt + \sigma\sqrt{z^2 - (X_t - m)^2}dW_t \tag{45}$$

where $\beta > 0$ and $\gamma \in (-1, 1)$ has been proposed as a model for the random variation of the logarithm of an exchange rate in a target zone between realignments by De Jong et al. (2001) ($\gamma = 0$) and Larsen and Sørensen (2007). This is a diffusion on the interval $(m - z, m + z)$ with mean reversion around $m + \gamma z$. The parameter γ quantifies the asymmetry of the model. When $\beta(1-\gamma) \geq \sigma^2$ and $\beta(1+\gamma) \geq \sigma^2$, X is an ergodic diffusion, for which the stationary distribution is a beta-distribution on $(m - z, m + z)$ with parameters $\kappa_1 = \beta(1 - \gamma)\sigma^{-2}$ and $\kappa_2 = \beta(1 + \gamma)\sigma^{-2}$.

The eigenfunctions for the generator of the diffusion (45) are

$$f_i(x;\beta,\gamma,\sigma,m,z) = P_i^{(\kappa_1 - 1,\, \kappa_2 - 1)}((x - m)/z), \quad i = 1, 2, \ldots$$

where $P_i^{(a,b)}(x)$ denotes the Jacobi polynomial of order i given by

$$P_i^{(a,b)}(x) = \sum_{j=0}^{i} 2^{-j} \binom{n+a}{n-j} \binom{a+b+n+j}{j} (x-1)^j, \quad -1 < x < 1,$$

as can easily be seem by direct calculation. For this reason, the process (45) is called a *Jacobi-diffusion*. The eigenvalue of f_i is $i\left(\beta + \frac{1}{2}\sigma^2(i-1)\right)$. Since condition (i) above is obviously satisfied because the state space is bounded, (44) holds. □

When the eigenfunctions are of the form

$$f_i(y;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\, \kappa(y)^j \tag{46}$$

where κ is a real function defined on the state space and is independent of θ, the optimal weight matrix (25) can be found explicitly too, provided that $2N$ eigenfunctions are available. Specifically,

$$B_h(x,\theta)_{ij} = \sum_{k=0}^{j} \left(\partial_{\theta_i} a_{j,k}(\theta) \nu_k(x;\theta) - \partial_{\theta_i} [e^{-\lambda_j(\theta)\Delta} \phi_j(x;\theta)] \right)$$

and

$$V_h(x,\theta)_{i,j} = \sum_{r=0}^{i} \sum_{s=0}^{j} \left(a_{i,r}(\theta) a_{j,s}(\theta) \nu_{r+s}(x;\theta) - e^{-[\lambda_i(\theta)+\lambda_j(\theta)]\Delta} \, \phi_i(x;\theta)\phi_j(x;\theta) \right),$$

where $\nu_i(x;\theta) = \pi_\Delta^\theta(\kappa^i)(x)$, $i = 1, \ldots, 2N$, solve the following triangular system of linear equations

$$e^{-\lambda_i(\theta)\Delta} f_i(x;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta) \nu_j(x;\theta) \quad i = 1, \ldots, 2N, \tag{47}$$

with $\nu_0(x;\theta) = 1$. The expressions for B_h and V_h follow from (34) and (35), while (47) follows by applying π_Δ^θ to both sides of (46).

Example 5 A widely applicable class of diffusion models for which explicit polynomial eigenfunctions are available is the class of Pearson diffusions, see Wong (1964) and Forman and Sørensen (2008). A Pearson diffusion is a stationary solution to a stochastic differential equation of the form

$$dX_t = -\beta(X_t - \mu)dt + \sqrt{(aX_t^2 + bX_t + c)} dW_t, \tag{48}$$

where $\beta > 0$, and a, b and c are such that the square root is well defined when X_t is in the state space. The class of stationary distributions equals the full Pearson system of distributions, so a very wide spectrum of marginal distributions is available ranging from distributions with compact support to very heavy-tailed distributions. For instance Pearson's type IV distributions, a skew t-type distribution, which seems very useful in finance, see e.g. Nagahara (1996), is the stationary distribution of the diffusion

$$dZ_t = -\beta Z_t dt + \sqrt{2\beta(\nu-1)^{-1}\{Z_t^2 + 2\rho\nu^{\frac{1}{2}} Z_t + (1+\rho^2)\nu\}} dW_t,$$

with $\nu > 1$. The parameter ρ is a skewness parameter. For $\rho = 0$ a t-distribution with ν degrees of freedom is obtained. Well-known instances of (48) are the Ornstein-Uhlenbeck process, the square root (CIR) process, and the Jacobi diffusions.

For a diffusion $T(X)$ obtained from a solution X to (48) by a twice differentiable and invertible transformation T, the eigenfunctions of the generator are $p_n\{T^{-1}(x)\}$, where p_n is an eigenfunction of the generator of X. The eigenvalues are the same as for the original eigenfunctions. Since the original eigenfunctions are polynomials, the eigenfunctions of $T(X)$ are of the form (46) with $\kappa = T^{-1}$. Hence explicit optimal martingale estimating functions are available for transformed Pearson diffusions too.

As an example let X be the Jacobi-diffusion (45) with $m = 0$ and $z = 1$, and consider $Y_t = \sin^{-1}(X_t)$. Then

$$dY_t = -\rho \frac{\sin(Y_t) - \varphi}{\cos(Y_t)} dt + \sigma dW_t,$$

where $\rho = \beta - \frac{1}{2}\sigma^2$ and $\varphi = \beta\gamma/\rho$. The state space is $(-\pi/2, \pi/2)$. Note that Y has dynamics that are very different from those of (45): the drift is non-linear and the diffusion coefficient is constant. The process Y was proposed and studied in Kessler and Sørensen (1999) for $\varphi = 0$, where the drift is $-\rho \tan(x)$. The general asymmetric version was proposed in Larsen and Sørensen (2007) as a model for exchange rates in a target zone. □

Explicit martingale estimating functions are only available for the relatively small, but versatile, class of diffusions for which explicit eigenfunctions for the generator are available. *Explicit non-martingale estimating functions* can be found for all diffusions, but cannot be expected to approximate the score functions as well as martingale estimating functions, and will therefore usually give less efficient estimators.

Hansen and Scheinkman (1995) proposed non-martingale estimating functions given by

$$g_j(\Delta, x, y; \theta) = h_j(y) A_\theta f_j(x) - f_j(x) \widehat{A}_\theta h_j(y), \tag{49}$$

where A_θ is the generator (38), and the functions f_j and h_j satisfy weak regularity conditions ensuring that (9) holds. The differential operator

$$\widehat{A}_\theta f(x) = \sum_{k=1}^{d} \widehat{b}_k(x;\theta) \partial_{x_k} f(x) + \frac{1}{2} \sum_{k,\ell=1}^{d} C_{k\ell}(x;\theta) \partial^2_{x_k x_\ell} f(x),$$

where $C = \sigma\sigma^T$ and

$$\widehat{b}_k(x;\theta) = -b_k(x;\theta) + \frac{1}{\mu_\theta(x)} \sum_{\ell=1}^{d} \partial_{x_\ell} \left(\mu_\theta C_{kl}\right)(x;\theta),$$

is the generator of the time reversal of the observed diffusion X. A simpler type of explicit non-martingale estimating functions is of the form $g(\Delta, x, y; \theta) = h(x; \theta)$. Hansen and Scheinkman (1995) and Kessler (2000) studied $h_j(x;\theta) = A_\theta f_j(x)$, which is a particular case of (49). Kessler

(2000) also proposed $h(x;\theta) = \partial_\theta \log \mu_\theta(x)$, which corresponds to considering the observations as an i.i.d. sample from the stationary distribution. Finally, Sørensen (2001) derived the estimating function with $h(x;\theta) = A_\theta \partial_\theta \log \mu_\theta(X_{t_i})$ as an approximation to the continuous-time score function. In all cases weak regularity conditions are needed to ensure that (9) holds, i.e. that $\int h(x;\theta_0)\mu_{\theta_0}(x)dx = 0$.

Quite generally, an explicit *approximate martingale estimating function* can be obtained from a martingale estimating function of the form (32) by approximating $\pi^\theta_\Delta(f_j(\theta))(x)$ and the weight matrix by (37). The simplest version of this approach gives the same estimator as the Gaussian quasi-likelihood based on the Euler-approximation to (1). Estimators of this type have been considered by Dorogovcev (1976), Prakasa Rao (1988), Florens-Zmirou (1989), Yoshida (1992), Chan et al. (1992), Kessler (1997), and Kelly et al. (2004). It is, however, important to note that there is a dangerous pitfall when using these simple approximate martingale estimating functions. They do not satisfy (9), and hence the estimators are inconsistent and converge to the solution to (12). The problem is illustrated by the following example.

Example 6 Consider again the CIR-model (42). If we insert the approximation $F(x;\alpha,\beta) = -\beta(x-\alpha)\Delta$ into (43) we obtain the following estimator for β

$$\widehat{\beta}_n = \frac{\frac{1}{n}(X_{\Delta n} - X_0)\sum_{i=1}^n X^{-1}_{\Delta(i-1)} - \sum_{i=1}^n X^{-1}_{\Delta(i-1)}(X_{\Delta i} - X_{\Delta(i-1)})}{\Delta[n - (\sum_{i=1}^n X_{\Delta(i-1)})(\sum_{i=1}^n X^{-1}_{\Delta(i-1)})/n]}.$$

It follows from (5) that

$$\widehat{\beta}_n \xrightarrow{P_\theta} (1 - e^{-\beta_0 \Delta})/\Delta \leq \Delta^{-1}.$$

Thus the estimator of the reversion parameter β is reasonable only when $\beta_0\Delta$ is considerably smaller than one. Note that the estimator will always converge to a limit smaller than the sampling frequency. When $\beta_0\Delta$ is large, the behaviour of the estimator is bizarre, see Bibby and Sørensen (1995). Without prior knowledge of the value of β_0 it is thus a dangerous estimator. □

The asymptotic bias given by (12) is small when Δ is sufficiently small, and the results in the following section on high frequency asymptotics show that in this case the approximate martingale estimating functions work well. However, how small Δ has to be depends on the parameter values, and without prior knowledge about the parameters, it is safer to use an exact martingale estimating function, which gives consistent estimators at all sampling frequencies.

6 High Frequency Asymptotics and Efficient Estimation

A large number of estimating functions have been proposed for diffusion models, and a large number of simulation studies have been performed to compare their relative merits, but the general picture has been rather confusing. By considering the high frequency scenario,

$$n \to \infty, \qquad \Delta_n \to 0, \qquad n\Delta_n \to \infty, \tag{50}$$

Sørensen (2007) obtained simple conditions for rate optimality and efficiency for ergodic diffusions, which allow identification of estimators that work well when the time between observations, Δ_n, is not too large. For financial data the speed of reversion is usually slow enough that this type of asymptotics works for daily, sometimes even weekly observations. A main result of this theory is that under weak conditions optimal martingale estimating functions give rate optimal and efficient estimators.

To simplify the exposition, we restrict attention to a one-dimensional diffusion given by

$$dX_t = b(X_t;\alpha)dt + \sigma(X_t;\beta)dW_t, \tag{51}$$

where $\theta = (\alpha, \beta) \in \Theta \subseteq \mathbb{R}^2$. The results below can be generalized to multivariate diffusions and parameters of higher dimension. We consider estimating functions of the general form (3), where the two-dimensional function $g = (g_1, g_2)$ for some $\kappa \geq 2$ and for all $\theta \in \Theta$ satisfies

$$E_\theta(g(\Delta_n, X_{\Delta_n i}, X_{\Delta_n(i-1)};\theta) \,|\, X_{\Delta_n(i-1)}) = \Delta_n^\kappa R(\Delta_n, X_{\Delta_n(i-1)};\theta). \tag{52}$$

Here and later $R(\Delta, y, x;\theta)$ denotes a function such that $|R(\Delta, y, x;\theta)| \leq F(y, x;\theta)$, where F is of polynomial growth in y and x uniformly for θ in a compact set[7]. We assume that the diffusion and the estimating functions satisfy the technical regularity Condition 6 given below.

Martingale estimating functions obviously satisfy (52) with $R = 0$, but for instance the approximate martingale estimating functions discussed at the end of the previous section satisfy (52) too.

Theorem 3 *Suppose that*

$$\partial_y g_2(0, x, x;\theta) = 0, \tag{53}$$

$$\partial_y g_1(0, x, x;\theta) = \partial_\alpha b(x;\alpha)/\sigma^2(x;\beta), \tag{54}$$

$$\partial_y^2 g_2(0, x, x;\theta) = \partial_\beta \sigma^2(x;\beta)/\sigma^2(x;\beta)^2, \tag{55}$$

for all $x \in (\ell, r)$ and $\theta \in \Theta$. Assume, moreover, that the following identifiability condition is satisfied

[7] For any compact subset $K \subseteq \Theta$, there exist constants $C_1, C_2, C_3 > 0$ such that $\sup_{\theta\in K} |F(y, x;\theta)| \leq C_1(1 + |x|_2^C + |y|_3^C)$ for all x and y in the state space of the diffusion.

$$\int_\ell^r [b(x,\alpha_0) - b(x,\alpha)]\partial_y g_1(0,x,x;\theta)\mu_{\theta_0}(x)dx \neq 0 \qquad \textit{when } \alpha \neq \alpha_0,$$

$$\int_\ell^r [\sigma^2(x,\beta_0) - \sigma^2(x,\beta)]\partial_y^2 g_2(0,x,x;\theta)\mu_{\theta_0}(x)dx \neq 0 \qquad \textit{when } \beta \neq \beta_0,$$

and that

$$W_1 = \int_\ell^r \frac{(\partial_\alpha b(x;\alpha_0))^2}{\sigma^2(x;\beta_0)}\mu_{\theta_0}(x)dx \neq 0,$$

$$W_2 = \int_\ell^r \left[\frac{\partial_\beta \sigma^2(x;\beta_0)}{\sigma^2(x;\beta_0)}\right]^2 \mu_{\theta_0}(x)dx \neq 0.$$

Then a consistent G_n–estimator $\widehat{\theta}_n = (\widehat{\alpha}_n, \widehat{\beta}_n)$ exists and is unique in any compact subset of Θ containing θ_0 with probability approaching one as $n \to \infty$. For a martingale estimating function, or more generally if $n\Delta^{2(\kappa-1)} \to 0$,

$$\begin{pmatrix} \sqrt{n\Delta_n}(\widehat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\widehat{\beta}_n - \beta_0) \end{pmatrix} \xrightarrow{\mathcal{D}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} W_1^{-1} & 0 \\ 0 & W_2^{-1} \end{pmatrix}\right). \tag{56}$$

An estimator satisfying (56) is rate optimal and efficient, cf. Gobet (2002), who showed that the model considered here is locally asymptotically normal. Note that the estimator of the diffusion coefficient parameter, β, converges faster than the estimator of the drift parameter, α. Condition (53) implies rate optimality. If this condition is not satisfied, the estimator of the diffusion coefficient parameter converges at the slower rate $\sqrt{n\Delta_n}$. This condition is called *the Jacobsen condition*, because it appears in the theory of *small Δ-optimal estimation* developed in Jacobsen (2001) and Jacobsen (2002). In this theory the asymptotic covariance matrix in (11) is expanded in powers of Δ, the time between observations. The leading term is minimal when (54) and (55) are satisfied. The same expansion of (11) was used by Aït-Sahalia and Mykland (2004).

The assumption $n\Delta_n \to \infty$ in (50) is needed to ensure that the drift parameter, α, can be consistently estimated. If the drift is known and only the diffusion coefficient parameter, β, needs to be estimated, this condition can be omitted, see Genon-Catalot and Jacod (1993). Another situation where the infinite observation horizon, $n\Delta_n \to \infty$, is not needed for consistent estimation of α is when the high frequency asymptotic scenario is combined with the small diffusion scenario, where $\sigma(x;\beta) = \epsilon_n \zeta(x;\beta)$ and $\epsilon_n \to 0$, see Genon-Catalot (1990), Sørensen and Uchida (2003) and Gloter and Sørensen (2008).

The reader is reminded of the trivial fact that for any non-singular 2×2 matrix, M_n, the estimating functions $M_n G_n(\theta)$ and $G_n(\theta)$ give exactly the same estimator. We call them *versions* of the same estimating function. The matrix M_n may depend on Δ_n. Therefore a given version of an estimating

function needs not satisfy (53) – (55). The point is that a version must exist which satisfies these conditions.

Example 7 Consider a quadratic martingale estimating function of the form

$$g(\Delta, y, x; \theta) = \begin{pmatrix} a_1(x, \Delta; \theta)[y - F(\Delta, x; \theta)] \\ a_2(x, \Delta; \theta)\left[(y - F(\Delta, x; \theta))^2 - \phi(\Delta, x; \theta)\right] \end{pmatrix}, \tag{57}$$

where F and ϕ are given by (19) and (20). By (37), $F(\Delta, x; \theta) = x + O(\Delta)$ and $\phi(\Delta, x; \theta) = O(\Delta)$, so

$$g(0, y, x; \theta) = \begin{pmatrix} a_1(x, 0; \theta)(y - x) \\ a_2(x, 0; \theta)(y - x)^2 \end{pmatrix}. \tag{58}$$

Since $\partial_y g_2(0, y, x; \theta) = 2a_2(x, \Delta; \theta)(y - x)$, the Jacobsen condition (53) is satisfied, so estimators obtained from (57) are rate optimal. Using again (37), it is not difficult to see that efficient estimators are obtained in three particular cases: the optimal estimating function given in Example 2 and the approximations (21) and (41). □

It follows from results in Jacobsen (2002) that to obtain a rate optimal and efficient estimator from an estimating function of the form (32), we need that $N \geq 2$ and that the matrix

$$D(x) = \begin{pmatrix} \partial_x f_1(x; \theta) & \partial_x^2 f_1(x; \theta) \\ \partial_x f_2(x; \theta) & \partial_x^2 f_2(x; \theta) \end{pmatrix}$$

is invertible for μ_θ-almost all x. Under these conditions, Sørensen (2007) showed that Godambe-Heyde optimal martingale estimating functions give rate optimal and efficient estimators. For a d-dimensional diffusion, Jacobsen (2002) gave the conditions $N \geq d(d+3)/2$, and that the $N \times (d + d^2)$-matrix $D(x) = \left(\partial_x f(x; \theta)\ \partial_x^2 f(x; \theta)\right)$ has full rank $d(d+3)/2$.

We conclude this section by stating technical conditions under which the results in this section hold. The assumptions about polynomial growth are far too strong, but simplify the proofs. These conditions can most likely be weakened very considerably in a way similar to the proofs in Gloter and Sørensen (2008).

convergence The diffusion is ergodic and the following conditions hold for all $\theta \in \Theta$:

(1)$\int_\ell^r x^k \mu_\theta(x) dx < \infty$ for all $k \in \mathbb{N}$.
(2)$\sup_t E_\theta(|X_t|^k) < \infty$ for all $k \in \mathbb{N}$.
(3)$b, \sigma \in C_{p,4,1}((\ell, r) \times \Theta)$.
(4)$g(\Delta, y, x; \theta) \in C_{p,2,6,2}(\mathbb{R}_+ \times (\ell, r)^2 \times \Theta)$ and has an expansion in powers of Δ:

$$g(\Delta, y, x; \theta) = g(0, y, x; \theta) + \Delta g^{(1)}(y, x; \theta) + \tfrac{1}{2}\Delta^2 g^{(2)}(y, x; \theta) + \Delta^3 R(\Delta, y, x; \theta),$$

where

$$g(0, y, x; \theta) \in C_{p,6,2}((\ell, r)^2 \times \Theta),$$
$$g^{(1)}(y, x; \theta) \in C_{p,4,2}((\ell, r)^2 \times \Theta),$$
$$g^{(2)}(y, x; \theta) \in C_{p,2,2}((\ell, r)^2 \times \Theta).$$

We define $C_{p,k_1,k_2,k_3}(\mathbb{R}_+ \times (\ell, r)^2 \times \Theta)$ as the class of real functions $f(t, y, x; \theta)$ satisfying that

(i) $f(t, y, x; \theta)$ is k_1 times continuously differentiable with respect t, k_2 times continuously differentiable with respect y, and k_3 times continuously differentiable with respect α and with respect to β

(ii) f and all partial derivatives $\partial_t^{i_1} \partial_y^{i_2} \partial_\alpha^{i_3} \partial_\beta^{i_4} f$, $i_j = 1, \ldots k_j$, $j = 1, 2$, $i_3 + i_4 \leq k_3$, are of polynomial growth in x and y uniformly for θ in a compact set (for fixed t).

The classes $C_{p,k_1,k_2}((\ell, r) \times \Theta)$ and $C_{p,k_1,k_2}((\ell, r)^2 \times \Theta)$ are defined similarly for functions $f(y; \theta)$ and $f(y, x; \theta)$, respectively.

References

Aït-Sahalia, Y. (2002): Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica* **70**, 223–262.

Aït-Sahalia, Y. (2003): Closed-form likelihood expansions for multivariate diffusions. *Working paper, Princeton University. Ann. Statist.* to appear.

Aït-Sahalia, Y. and Mykland, P. (2003): The effects of random and discrete sampling when estimating continuous-time diffusions. *Econometrica* **71**, 483–549.

Aït-Sahalia, Y. and Mykland, P. A. (2004): Estimators of diffusions with randomly spaced discrete observations: a general theory. *Ann. Statist.* **32**, 2186–2222.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006): Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. Roy. Statist. Soc. B* **68**, 333–382.

Bibby, B. M. and Sørensen, M. (1995): Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1**, 17–39.

Bibby, B. M. and Sørensen, M. (1996): On estimation for discretely observed diffusions: a review. *Theory of Stochastic Processes* **2**, 49–56.

Billingsley, P. (1961): The lindeberg-lévy theorem for martingales. *Proc. Amer. Math. Soc.* **12**, 788–792.

Bollerslev, T. and Wooldridge, J. (1992): Quasi-maximum likelihood estimators and inference in dynamic models with time-varying covariances. *Econometric Review* **11**, 143–172.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997): *The Econometrics of Financial Markets*. Princeton University Press, Princeton.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992): An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* **47**, 1209–1227.

Dacunha-Castelle, D. and Florens-Zmirou, D. (1986): Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19**, 263–284.

De Jong, F., Drost, F. C., and Werker, B. J. M. (2001): A jump-diffusion model for exchange rates in a target zone. *Statistica Neerlandica* **55**, 270–300.

Dorogovcev, A. J. (1976): The consistency of an estimate of a parameter of a stochastic differential equation. *Theor. Probability and Math. Statist.* **10**, 73–82.

Doukhan, P. (1994): *Mixing, Properties and Examples. Lecture Notes in Statistics 85.* Springer, New York.

Durbin, J. (1960): Estimation of parameters in time-series regression models. *J. Roy. Statist. Soc. B* **22**, 139–153.

Durham, G. B. and Gallant, A. R. (2002): Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *J. Business & Econom. Statist.* **20**, 297–338.

Elerian, O., Chib, S., and Shephard, N. (2001): Likelihood inference for discretely observed non-linear diffusions. *Econometrica* **69**, 959–993.

Eraker, B. (2001): Mcmc analysis of diffusion models with application to finance. *J. Business & Econom. Statist.* **19**, 177–191.

Florens-Zmirou, D. (1989): Approximate discrete-time schemes for statistics of diffusion processes. *Statistics* **20**, 547–557.

Forman, J. L. and Sørensen, M. (2008): The pearson diffusions: A class of statistically tractable diffusion processes. *Scand. J. Statist.* to appear.

Genon-Catalot, V. (1990): Maximum contrast estimation for diffusion processes from discrete observations. *Statistics* **21**, 99–116.

Genon-Catalot, V. and Jacod, J. (1993): On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. Henri Poincaré, Probabilités et Statistiques* **29**, 119–151.

Genon-Catalot, V., Jeantheau, T., and Larédo, C. (2000): Stochastic volatility models as hidden markov models and statistical applications. *Bernoulli* **6**, 1051–1079.

Gloter, A. and Sørensen, M. (2008): Estimation for stochastic differential equations with a small diffusion coefficient. *Stoch. Proc. Appl.* to appear.

Gobet, E. (2002): Lan property for ergodic diffusions with discrete observations. *Ann. Inst. Henri Poincaré, Probabilités et Statistiques* **38**, 711–737.

Godambe, V. P. (1960): An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.* **31**, 1208–1212.

Godambe, V. P. and Heyde, C. C. (1987): Quasi likelihood and optimal estimation. *International Statistical Review* **55**, 231–244.

Hall, P. and Heyde, C. C. (1980): *Martingale Limit Theory and Its Applications.* Academic Press, New York.

Hansen, L. P. (1982): Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Hansen, L. P. and Scheinkman, J. A. (1995): Back to the future: generating moment implications for continuous-time markov processes. *Econometrica* **63**, 767–804.

Heyde, C. C. (1997): *Quasi-Likelihood and Its Application.* Springer-Verlag, New York.

Jacobsen, M. (2001): Discretely observed diffusions; classes of estimating functions and small δ-optimality. *Scand. J. Statist.* **28**, 123–150.

Jacobsen, M. (2002): Optimality and small δ-optimality of martingale estimating functions. *Bernoulli* **8**, 643–668.

Jacod, J. and Sørensen, M. (2008): Asymptotic statistical theory for stochastic processes: a review. *Preprint, Department of Mathematical Sciences, University of Copenhagen.*

Kelly, L., Platen, E., and Sørensen, M. (2004): Estimation for discretely observed diffusions using transform functions. *J. Appl. Prob.* **41**, 99–118.

Kessler, M. (1996): *Estimation paramétrique des coefficients d'une diffusion ergodique à partir d'observations discrètes*. PhD thesis, Laboratoire de Probabilités, Université Paris VI.

Kessler, M. (1997): Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24**, 211–229.

Kessler, M. (2000): Simple and explicit estimating functions for a discretely observed diffusion process. *Scand. J. Statist.* **27**, 65–82.

Kessler, M. and Paredes, S. (2002): Computational aspects related to martingale estimating functions for a discretely observed diffusion. *Scand. J. Statist.* **29**, 425–440.

Kessler, M. and Sørensen, M. (1999): Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli* **5**, 299–314.

Kusuoka, S. and Yoshida, N. (2000): Malliavin calculus, geometric mixing, and expansion of diffusion functionals. *Probability Theory and Related Fields* **116**, 457–484.

Larsen, K. S. and Sørensen, M. (2007): A diffusion model for exchange rates in a target zone. *Mathematical Finance* **17**, 285–306.

Nagahara, Y. (1996): Non-gaussian distribution for stock returns and related stochastic differential equation. *Financial Engineering and the Japanese Markets* **3**, 121–149.

Overbeck, L. and Rydén, T. (1997): Estimation in the cox-ingersoll-ross model. *Econometric Theory* **13**, 430–461.

Ozaki, T. (1985): Non-linear time series models and dynamical systems. *In: Hannan, E. J., Krishnaiah, P. R., and Rao, M. M. (Eds.): Handbook of Statistics* **5**, 25–83. Elsevier Science Publishers.

Pedersen, A. R. (1995): A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.* **22**, 55–71.

Phillips, P.C.B. and Yu, J. (2008): Maximum likelihood and Gaussian estimation of continuous time models in finance. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*. 497–530. Springer, New York.

Poulsen, R. (1999): Approximate maximum likelihood estimation of discretely observed diffusion processes. *Working Paper 29, Centre for Analytical Finance, Aarhus*.

Prakasa Rao, B. L. S. (1988): Statistical inference from sampled data for stochastic processes. *Contemporary Mathematics* **80**, 249–284.

Roberts, G. O. and Stramer, O. (2001): On inference for partially observed nonlinear diffusion models using metropolis-hastings algorithms. *Biometrika* **88**, 603–621.

Skorokhod, A. V. (1989): *Asymptotic Methods in the Theory of Stochastic Differential Equations*. American Mathematical Society, Providence, Rhode Island.

Sørensen, H. (2001): Discretely observed diffusions: Approximation of the continuous-time score function. *Scand. J. Statist.* **28**, 113–121.

Sørensen, M. (1997): Estimating functions for discretely observed diffusions: a review. *In: Basawa, I. V., Godambe, V. P. and Taylor, R. L. (Eds.): Selected Proceedings of the Symposium on Estimating Functions*, 305–325. IMS Lecture Notes–Monograph Series **32**. Institute of Mathematical Statistics, Hayward.

Sørensen, M. (2007): Efficient estimation for ergodic diffusions sampled at high frequency. *Preprint, Department of Mathematical Sciences, University of Copenhagen*.

Sørensen, M. and Uchida, M. (2003): Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli* **9**, 1051–1069.

Veretennikov, A. Y. (1987): Bounds for the mixing rate in the theory of stochastic equations. *Theory of Probability and its Applications* **32**, 273–281.

Wong, E. (1964): The construction of a class of stationary markoff processes. *In: Bellman, R. (Ed.): Stochastic Processes in Mathematical Physics and Engineering*, 264–276. American Mathematical Society, Rhode Island.

Yoshida, N. (1992): Estimation for diffusion processes from discrete observations. *Journal of Multivariate Analysis* **41**, 220–242.

Realized Volatility

Torben G. Andersen and Luca Benzoni *

Abstract Realized volatility is a nonparametric ex-post estimate of the return variation. The most obvious realized volatility measure is the sum of finely-sampled squared return realizations over a fixed time interval. In a frictionless market the estimate achieves consistency for the underlying quadratic return variation when returns are sampled at increasingly higher frequency. We begin with an account of how and why the procedure works in a simplified setting and then extend the discussion to a more general framework. Along the way we clarify how the realized volatility and quadratic return variation relate to the more commonly applied concept of conditional return variance. We then review a set of related and useful notions of return variation along with practical measurement issues (e.g., discretization error and microstructure noise) before briefly touching on the existing empirical applications.

Torben G. Andersen
Kellogg School of Management, Northwestern University, Evanston, IL; NBER, Cambridge, MA; and CREATES, Aarhus, Denmark, e-mail: t-andersen@northwestern.edu

Luca Benzoni
Federal Reserve Bank of Chicago, Chicago, IL, e-mail: lbenzoni@frbchi.org

* We are grateful to Neil Shephard, Olena Chyruk, and the Editors Richard Davis and Thomas Mikosch for helpful comments and suggestions. Of course, all errors remain our sole responsibility. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. The work of Andersen is supported by a grant from the NSF to the NBER and support from CREATES funded by the Danish National Research Foundation.

DOI: 10.1007/978-3-540-71297-8_24,

1 Introduction

Given the importance of return volatility on a number of practical financial management decisions, there have been extensive efforts to provide good real-time estimates and forecasts of current and future volatility. One complicating feature is that, contrary to the raw return, actual realizations of return volatility are not directly observable. A common approach to deal with the fundamental latency of return volatility is to conduct inference regarding volatility through strong parametric assumptions, invoking, e.g., an ARCH or a stochastic volatility (SV) model estimated with data at daily or lower frequency. An alternative approach is to invoke option pricing models to invert observed derivatives prices into market-based forecasts of "implied volatility" over a fixed future horizon. Such procedures remain model-dependent and further incorporate a potentially time-varying volatility risk premium in the measure so they generally do not provide unbiased forecasts of the volatility of the underlying asset. Finally, some studies rely on "historical" volatility measures that employ a backward looking rolling sample return standard deviation, typically computed using one to six months of daily returns, as a proxy for the current and future volatility level. Since volatility is persistent such measures do provide information but volatility is also clearly mean reverting, implying that such unit root type forecasts of future volatility are far from optimal and, in fact, conditionally biased given the history of the past returns. In sum, while actual returns may be measured with minimal (measurement) error and may be analyzed directly via standard time series methods, volatility modeling has traditionally relied on more complex econometric procedures in order to accommodate the inherent latent character of volatility.

The notion of realized volatility effectively reverses the above characterization. Given continuously observed price or quote data, and absent transaction costs, the realized return variation may be measured without error along with the (realized) return. In addition, the realized variation is conceptually related to the cumulative expected variability of the returns over the given horizon for a wide range of underlying arbitrage-free diffusive data generating processes. In contrast, it is impossible to relate the actual (realized) return to the expected return over shorter sample periods in any formal manner absent very strong auxiliary assumptions. In other words, we learn much about the expected return volatility and almost nothing about the expected mean return from finely-sampled asset prices. This insight has fueled a dramatic increase in research into the measurement and application of realized volatility measures obtained from high frequency, yet noisy, observations on returns. For liquid financial markets with high trade and quote frequency and low transaction costs, it is now prevailing practice to rely on intra-day return data to construct ex-post volatility measures. Given the rapidly increasing availability of high-quality transaction data across many financial assets, it

is inevitable that this approach will continue to be developed and applied within ever broader contexts in the future.

This chapter provides a short and largely intuitive overview of the realized volatility concept and the associated applications. We begin with an account of how and why the procedure works in a simplified setting and then discuss more formally how the results apply in general settings. Next, we detail more formally how the realized volatility and quadratic return variation relate to the more common conditional return variance concept. We then review a set of related and useful notions of return variation along with practical measurement issues before briefly touching on the existing empirical applications.

2 Measuring Mean Return versus Return Volatility

The theory of realized volatility is tied closely to the availability of asset price observations at arbitrarily high frequencies. Hence, it is natural to consider the volatility measurement problem in a continuous-time framework, even if we ultimately only allow sampling at discrete intervals. We concentrate on a single risky asset whose price may be observed at equally-spaced discrete points in time over a given interval, $[0, T]$, namely $t = 0, 1/n, 2/n, \ldots, T - (1/n), T$, where n and T are positive integers and the unit interval corresponds to the primary time period over which we desire to measure return volatility, e.g., one trading day. We denote the logarithmic asset price at time t by $s(t)$ and the continuously compounded returns over $[t-k, t]$ is then given by $r(t, k) = s(t) - s(t-k)$ where $0 \leq t - k < t \leq T$ and $k = j/n$ for some positive integer j. When $k = 1$ it is convenient to use the shorthand notation $r(t) = r(t, 1)$, where t is an integer $1 \leq t \leq T$, for the unit period, or "daily," return.

To convey the basic rationale behind the realized volatility approach, we initially consider a simplified setting with the continuously compounded returns driven by a simple time-invariant Brownian motion, so that

$$ds(t) = \alpha dt + \sigma dW(t),\ 0 \leq t \leq T\,, \tag{1}$$

where α and σ ($\sigma > 0$) denote the constant drift and diffusion coefficients, respectively, scaled to correspond to the unit time interval.

For a given measurement period, say $[0, K]$, where K is a positive integer, we have $n \cdot K$ intraday return observations $r(t, 1/n) = s(t) - s(t - 1/n)$ for $t = 1/n, \ldots, (n-1) \cdot K/n, K$, that are i.i.d. normally distributed with mean α/n and variance σ^2/n. It follows that the maximum likelihood estimator for the drift coefficient is given by

$$\widehat{\alpha}_n = \frac{1}{K} \sum_{j=1}^{n \cdot K} r(j/n, 1/n) = \frac{r(K, K)}{K} = \frac{s(K) - s(0)}{K}. \tag{2}$$

Hence, for a fixed interval the in-fill asymptotics, obtained by continually increasing the number of intraday observations, are irrelevant for estimating the expected return. The estimator of the drift is independent of the sampling frequency, given by n, and depends only on the span of the data, K. For example, one may readily deduce that

$$\mathrm{Var}(\widehat{\alpha}_n) = \frac{\sigma^2}{K}\,. \tag{3}$$

In other words, although the estimator is unbiased, the mean drift cannot be estimated consistently over any fixed interval. Even for the simplest case of a constant mean, long samples (large K) are necessary for precise inference. Thus, in a setting where the expected returns are stipulated to vary conditionally on features of the underlying economic environment, auxiliary identifying assumptions are required for sensible inference about α. This is the reason why critical empirical questions such as the size of the equity premium and the pattern of the expected returns in the cross-section of individual stocks remain contentious and unsettled issues within financial economics.

The situation is radically different for estimation of return volatility. Even if the expected return cannot be inferred with precision, nonparametric measurement of volatility may be based on un-adjusted or un-centered squared returns. This is feasible as the second return moment dominates the first moment in terms of influencing the high-frequency squared returns. Specifically, we have,

$$\mathrm{E}\big[\, r(j/n, 1/n)^2 \,\big] = \frac{\alpha^2}{n^2} + \frac{\sigma^2}{n}\,, \tag{4}$$

and

$$\mathrm{E}\big[\, r(j/n, 1/n)^4 \,\big] = \frac{\alpha^4}{n^4} + 6\frac{\alpha^2\sigma^2}{n^3} + 3\frac{\sigma^4}{n^2}\,. \tag{5}$$

It is evident that the terms involving the drift coefficient are an order of magnitude smaller, for n large, than those that pertain only to the diffusion coefficient. This feature allows us to estimate the return variation with a high degree of precision even without specifying the underlying mean drift, e.g.,[1]

$$\widehat{\sigma}^2_n = \frac{1}{K}\sum_{j=1}^{n\cdot K} r^2(j/n, 1/n)\,. \tag{6}$$

It is straightforward to establish that

$$\mathrm{E}\big[\, \widehat{\sigma}^2_n \,\big] = \frac{\alpha^2}{n} + \sigma^2\,, \tag{7}$$

while some additional calculations yield

[1] The quantity $(K \cdot \widehat{\sigma}^2_n)$ is a "realized volatility" estimator of the return variation over $[0, K]$ and it moves to the forefront of our discussion in the following section.

$$\mathrm{Var}\left[\,\widehat{\sigma}_n^2\,\right] = 4\frac{\alpha^2\sigma^2}{n^2K} + 2\frac{\sigma^4}{nK}\,. \tag{8}$$

It follows by a standard L^2 argument that, in probability, $\widehat{\sigma}_n^2 \to \sigma^2$ for $n \to \infty$. Hence, the realized variation measure is a biased but consistent estimator of the underlying (squared) volatility coefficient. Moreover, it is evident that, for n large, the bias is close to negligible. In fact, as $n \to \infty$ we have the distributional convergence,

$$\sqrt{n \cdot K}\,(\widehat{\sigma}_n^2 - \sigma^2) \;\to\; N(0, 2\sigma^4)\,. \tag{9}$$

These insights are not new. For example, within a similar context, they were stressed by Merton (1980). However, the lack of quality intraday price data and the highly restrictive setting have long led scholars to view them as bereft of practical import. This situation has changed fundamentally over the last decade, as it has been shown that the basic results apply very generally, high-frequency data have become commonplace, and the measurement procedures, through suitable strategies, can be adapted to deal with intraday observations for which the relative impact of microstructure noise may be substantial.

3 Quadratic Return Variation and Realized Volatility

This section outlines the main steps in generalizing the above findings to an empirically relevant setting with stochastic volatility. We still operate within the continuous-time diffusive setting, for simplicity ruling out price jumps, and assume a *frictionless market*. In this setting the asset's logarithmic price process s must be a semimartingale to rule out arbitrage opportunities (e.g., Back (1991)). We then have,

$$ds(t) = \mu(t)dt + \sigma(t)\,dW(t)\,, \qquad 0 \le t \le T\,, \tag{10}$$

where W is a standard Brownian motion process, $\mu(t)$ and $\sigma(t)$ are predictable processes, $\mu(t)$ is of finite variation, while $\sigma(t)$ is strictly positive and square integrable, i.e., $\mathrm{E}\left(\int_0^t \sigma_s^2 \mathrm{d}s\right) < \infty$. Hence, the processes $\mu(t)$ and $\sigma(t)$ signify the instantaneous conditional mean and volatility of the return. The continuously compounded return over the time interval from $t-k$ to t, $0 < k \le t$, is therefore

$$r(t,k) = s(t) - s(t-k) = \int_{t-k}^{t} \mu(\tau)d\tau + \int_{t-k}^{t} \sigma(\tau)dW(\tau)\,, \tag{11}$$

and its *quadratic variation* $QV(t,k)$ is

$$QV(t,k) = \int_{t-k}^{t} \sigma^2(\tau) d\tau \,. \tag{12}$$

Equation (12) shows that innovations to the mean component $\mu(t)$ do not affect the sample path variation of the return. Intuitively, this is because the mean term, $\mu(t)dt$, is of lower order in terms of second order properties than the diffusive innovations, $\sigma(t)dW(t)$. Thus, when cumulated across many high-frequency returns over a short time interval of length k they can effectively be neglected. The *diffusive* sample path variation over $[t-k,t]$ is also known as the *integrated variance* $IV(t,k)$,

$$IV(t,k) = \int_{t-k}^{t} \sigma^2(\tau) d\tau \,. \tag{13}$$

Equations (12) and (13) show that, in this setting, the quadratic and integrated variation coincide. This is however no longer true for more general return process like, e.g., the stochastic volatility jump-diffusion model discussed in Section 5 below.

Absent microstructure noise and measurement error, the return quadratic variation can be approximated arbitrarily well by the corresponding cumulative squared return process. Consider a partition $\{t-k+\frac{j}{n}, j=1,\ldots n\cdot k\}$ of the $[\,t-k,\,t\,]$ interval. Then the realized volatility (RV) of the logarithmic price process is

$$RV(t,k;n) = \sum_{j=1}^{n\cdot k} r\left(t-k+\frac{j}{n}, \frac{1}{n}\right)^2 . \tag{14}$$

Semimartingale theory ensures that the realized volatility measure converges in probability to the return quadratic variation QV, previously defined in equation (12), when the sampling frequency n increases:

$$RV(t,k;n) \longrightarrow QV(t,k) \quad \text{as } n \to \infty \,. \tag{15}$$

This finding extends the consistency result for the (constant) volatility coefficient discussed below equation (8) to a full-fledged stochastic volatility setting. This formal link between realized volatility measures based on high-frequency returns and the quadratic variation of the underlying (no arbitrage) price process follows immediately from the theory of semimartingales (e.g., Protter (1990)) and was first applied in the context of empirical return volatility measurement by Andersen and Bollerslev (1998a). The distributional result in equation (9) also generalizes directly, as we have, for $n \to \infty$,

$$\sqrt{n\cdot k}\left(\frac{RV(t,k;n)-QV(t,k)}{\sqrt{2\,IQ(t,k)}}\right) \to N(0,1)\,, \tag{16}$$

where $IQ(t,k) \equiv \int_{t-k}^{t} \sigma^4(\tau)d\tau$ is the integrated quarticity, with $IQ(t,k)$ independent from the limiting Gaussian distribution on the right hand side. This result was developed and brought into the realized volatility literature by Barndorff-Nielsen and Shephard (2002).[2]

Equation (16) sets the stage for formal ex-post inference regarding the actual realized return variation over a given period. However, the result is not directly applicable as the so-called integrated quarticity, $IQ(t,k)$, is unobserved and is likely to display large period-to-period variation. Hence, a consistent estimator for the integrated quarticity must be used in lieu of the true realization to enable feasible inference. Such estimators, applicable for any integrated power of the diffusive coefficient, have been proposed by Barndorff-Nielsen and Shephard (2002). The realized power variation of order p, $V(p;t,k;n)$ is the (scaled) cumulative sum of the absolute p-th power of the high-frequency returns and it converges, as $n \to \infty$, to the corresponding power variation of order p, $V(p;t,k)$. That is, defining the p-th realized power variation as,

$$V(p;t,k;n) \equiv n^{p/2-1}\,\mu_p^{-1} \sum_{j=1}^{n\cdot k} \left| r\left(t-k+\frac{j}{n},\frac{1}{n}\right)\right|^p, \tag{17}$$

where μ_p denotes the p-th absolute moment of a standard normal variable, we have, in probability,

$$V(p;t,k;n) \to \int_{t-k}^{t} \sigma^p(\tau)d\tau \equiv V(p;t,k)\,. \tag{18}$$

In other words, $V(4;t,k;n)$ is a natural choice as a consistent estimator for the integrated quarticity $IQ(t,k)$. It should be noted that this conclusion is heavily dependent on the absence of jumps in the price process which is an issue we address in more detail later. Moreover, the notion of realized power variation is a direct extension of realized volatility as $RV(t,k;n) = V(2;t,k;n)$ so equation (18) reduces to equation (15) for $p = 2$.

More details regarding the asymptotic results and multivariate generalizations of realized volatility may be found in, e.g., Andersen et al. (2001, 2003), Barndorff-Nielsen and Shephard (2001, 2002, 2004a), Meddahi (2002a), and Mykland (2006).

4 Conditional Return Variance and Realized Volatility

This section discusses the relationship between quadratic variation or integrated variance along with its associated empirical measure, realized volatil-

[2] The unpublished note by Jacod (1992) implies the identical result but this note was not known to the literature at the time.

ity, and the conditional return variance. In the case of constant drift and volatility coefficients, the conditional (and unconditional) return variance equals the quadratic variation of the log price process. In contrast, when volatility is stochastic we must distinguish clearly between the conditional variance, representing the (ex-ante) expected size of future squared return innovations over a certain period, and the quadratic variation, reflecting the actual (ex-post) realization of return variation, over the corresponding horizon. Hence, the distinction is one of a priori expectations versus subsequent actual realizations of return volatility. Under ideal conditions, the realized volatility captures the latter, but not the former. Nonetheless, realized volatility measures are useful in gauging the conditional return variance as one may construct well calibrated forecasts (conditional expectations) of return volatility from a time series of past realized volatilities. In fact, within a slightly simplified setting, we can formally strengthen these statements. If the instantaneous return is the continuous-time process (10) *and* the return, mean, and volatility processes are uncorrelated (i.e., $dW(t)$ and innovations to $\mu(t)$ and $\sigma(t)$ are mutually independent), then $r(t,k)$ is normally distributed conditional on the cumulative drift $\mu(t,k) \equiv \int_{t-k}^{t} \mu(\tau)d\tau$ and the quadratic variation $QV(t,k)$ (which in this setting equals the integrated variance $IV(t,k)$ as noted in equations (12) and (13)):

$$\big(r(t,k) \,|\, \mu(t,k), IV(t,k)\big) \;\sim\; \mathrm{N}\big(\,\mu(t,k), IV(t,k)\,\big)\,. \tag{19}$$

Consequently, the return distribution is mixed Gaussian with the mixture governed by the realizations of the integrated variance (and integrated mean) process. Extreme realizations (draws) from the integrated variance process render return outliers likely while persistence in the integrated variance process induces volatility clustering. Moreover, for short horizons, where the conditional mean is negligible relative to the cumulative absolute return innovations, the integrated variance may be directly related to the conditional variance as,

$$\mathrm{Var}[\,r(t,k) \,|\, \mathcal{F}_{t-k}\,] \approx \mathrm{E}[\,RV(t,k;n) \,|\, \mathcal{F}_{t-k}\,] \approx \mathrm{E}[\,QV(t,k) \,|\, \mathcal{F}_{t-k}\,]\,. \tag{20}$$

A volatility forecast is an estimate of the conditional return variance on the far left-hand side of equation (20), which in turn approximates the expected quadratic variation. Since RV is approximately unbiased for the corresponding unobserved quadratic variation, the realized volatility measure is the natural benchmark against which to gauge the performance of volatility forecasts. Goodness-of-fit tests may be conducted on the residuals given by the difference between the ex-post realized volatility measure and the ex-ante forecast. We review some of the evidence obtained via applications inspired by these relations in Section 7. In summary, the quadratic variation is directly related to the *actual* return variance as demonstrated by equation (19) and to the *expected* return variance, as follows from equation (20).

Finally, note that the realized volatility concept is associated with the return variation measured over a discrete time interval rather than with the so-called spot or instantaneous volatility. This distinction separates the realized volatility approach from a voluminous literature in statistics seeking to estimate spot volatility from discrete observations, predominantly in a setting with a constant diffusion coefficient. It also renders it distinct from the early contributions in financial econometrics allowing explicitly for time-varying volatilities, e.g., Foster and Nelson (1996). In principle, the realized volatility measurement can be adapted to spot volatility estimation: as k goes to zero, $QV(t,k)$ converges to the instantaneous volatility $\sigma^2(t)$, i.e., in principle RV converges to instantaneous volatility when both k and k/n shrink. For this to happen, however, k/n must converge at a rate higher than k, so as the interval shrinks we must sample returns at an ever increasing frequency. In practice, this is infeasible, because intensive sampling over tiny intervals magnifies the effects of microstructure noise. We return to this point in Section 6 where we discuss the bias in RV measures when returns are sampled with error.

5 Jumps and Bipower Variation

The return process in equation (10) is continuous under the stated regularity conditions, even if σ may display jumps. This is quite restrictive as asset prices often appear to exhibit sudden discrete movements when unexpected news hits the market. A broad class of SV models that allow for the presence of jumps in returns is defined by

$$ds(t) = \mu(t)dt + \sigma(t)dW(t) + \xi(t)\, dq_t\,, \tag{21}$$

where q is a Poisson process uncorrelated with W and governed by the jump intensity λ_t, i.e., $\text{Prob}(dq_t = 1) = \lambda_t\, dt$, with λ_t positive and finite. This assumption implies that there can only be a finite number of jumps in the price path per time period. This is a common restriction in the finance literature, though it rules out infinite activity Lévy processes. The scaling factor $\xi(t)$ denotes the magnitude of the jump in the return process if a jump occurs at time t. While explicit distributional assumptions often are invoked for parametric estimation, such restrictions are not required as the realized volatility approach is fully nonparametric in this dimension as well.

In this case, the quadratic return variation process over the interval from $t-k$ to t, $0 \le k \le t \le T$, is the sum of the diffusive integrated variance and the cumulative squared jumps:

$$QV(t,k) = \int_{t-k}^{t} \sigma^2(s)ds + \sum_{t-k\le s\le t} J^2(s) \equiv IV(t,k) + \sum_{t-k\le s\le t} J^2(s)\,, \tag{22}$$

where $J(t) \equiv \xi(t)dq(t)$ is non-zero only if there is a jump at time t.

The RV estimator (14) remains a consistent measure of the total QV in the presence of jumps, i.e., result (15) still holds; see, e.g., Protter (1990) and the discussion in Andersen et al. (2004). However, since the diffusive and jump volatility components appear to have distinctly different persistence properties it is useful both for analytic and predictive purposes to obtain separate estimates of these two factors in the decomposition of the quadratic variation implied by equation (22).

To this end, the h-skip bipower variation, BV, introduced by Barndorff-Nielsen and Shephard (2004b) provides a consistent estimate of the IV component,

$$BV(t,k;h,n) = \frac{\pi}{2} \sum_{i=h+1}^{n \cdot k} \left| r\left(t-k+\frac{ik}{n},\frac{1}{n}\right) \right| \left| r\left(t-k+\frac{(i-h)k}{n},\frac{1}{n}\right) \right|. \tag{23}$$

Setting $h = 1$ in definition (23) yields the 'realized bipower variation' $BV(t,k;n) \equiv BV(t,k;1,n)$. The bipower variation is robust to the presence of jumps and therefore, in combination with RV, it yields a consistent estimate of the cumulative squared jump component:

$$RV(t,k;n) - BV(t,k;n) \underset{n\to\infty}{\longrightarrow} QV(t,k) - IV(t,k) = \sum_{t-k\le s\le t} J^2(s)\,. \tag{24}$$

The results in equations (22)-(24) along with the associated asymptotic distributions have been exploited to improve the accuracy of volatility forecasts and to design tests for the presence of jumps in volatility. We discuss these applications in Section 7 below.

6 Efficient Sampling versus Microstructure Noise

The convergence relation in equation (15) states that RV approximates QV arbitrarily well as the sampling frequency n increases. Two issues, however, complicate the application of this result. First, even for the most liquid assets a continuous price record is unavailable. This limitation introduces an inevitable discretization error in the RV measures which forces us to recognize the presence of a measurement error. Although we may gauge the magnitude of such errors via the continuous record asymptotic theory outlined in equations (16)-(18), such inference is always subject to some finite sample distortions and it is only strictly valid in the absence of price jumps. Second, a wide array of microstructure effects induces spurious autocorrelations in the ultra-high frequency return series. The list includes price discreteness and rounding, bid-ask bounces, trades taking places on different markets and networks, gradual response of prices to a block trade, difference in information

contained in order of different size, strategic order flows, spread positioning due to dealer inventory control, and, finally, data recording mistakes. Such "spurious" autocorrelations can inflate the RV measures and thus generate a traditional type of bias-variance trade off. The highest possible sampling frequency should be used for efficiency. However, sampling at ultra-high frequency tends to bias the RV estimate.

A useful tool to assess this trade-off is the *volatility signature plot*, which depicts the sample average of the RV estimator over a long time span as a function of the sampling frequency. The long time span mitigates the impact of sampling variability so, absent microstructure noise, the plot should be close to a horizontal line. In practice, however, for transaction data obtained from liquid stocks the plot spikes at high sampling frequencies and decays rather smoothly to stabilize at frequencies in the 5- to 40-minute range. In contrast, the opposite often occurs for returns constructed from bid-ask quote midpoints as asymmetric adjustment of the spread induces positive serial correlation and biases the signature plot downward at the very highest sampling frequencies. Likewise, for illiquid stocks the inactive trading induces positive return serial autocorrelation which renders the signature plot increasing at lower sampling frequencies, see, e.g., Andersen et al. (2000a). Aït-Sahalia et al. (2005) and Bandi and Russell (2007) extend this approach by explicitly trading off efficient sampling versus bias-inducing noise to derive optimal sampling schemes.

Other researchers have suggested dealing with the problem by using alternative QV estimators that are less sensitive to microstructure noise. For instance, Huang and Tauchen (2005) and Andersen et al. (2007) note that using staggered returns and BV helps reduce the effect of noise, while Andersen et al. (2006a) extend volatility signature plots to include power and h-skip bipower variation. Other studies have instead relied on the high-low price range estimator (e.g., Alizadeh et al. (2002), Brandt and Diebold (2006), Brandt and Jones (2006), Gallant et al. (1999), Garman and Klass (1980), Parkinson (1980), Schwert (1990), and Yang and Zhang (2000)) to deal with situations in which the noise to signal ratio is high. Christensen and Podolskij (2006) and Dobrev (2007) generalize the range estimator to high-frequency data in distinct ways and discuss the link to RV.

A different solution to the problem is considered in the original contribution of Zhou (1996) who seeks to correct the bias of RV style estimators by explicitly accounting for the covariance in lagged squared return observations. Hansen and Lunde (2006) extend Zhou's approach to the case of non-i.i.d. noise. In contrast, Aït-Sahalia et al. (2005) explicitly determine the requisite bias correction when the noise term is i.i.d. normally distributed, while Zhang et al. (2005) propose a consistent volatility estimator that uses the entire price record by averaging RVs computed from different sparse subsamples and correcting for the remaining bias. Aït-Sahalia et al. (2006) extend the sub-sampling approach to account for certain types of serially correlated

errors. Another prominent and general approach is the recently proposed kernel-based technique of Barndorff-Nielsen et al. (2006a, 2006b).

7 Empirical Applications

Since the early 1990s transaction data have become increasingly available to academic research. This development has opened the way for a wide array of empirical applications exploiting the realized return variation approach. Below we briefly review the progress in different areas of research.

7.1 Early work

Hsieh (1991) provides one of the first estimates of the daily return variation constructed from intra-daily S&P500 returns sampled at the 15-minute frequency. The investigation is informal in the sense that there is no direct association with the concept of quadratic variation. More in-depth applications were pursued in publications by the Olsen & Associates group and later surveyed in Dacorogna et al. (2001) as they explore both intraday periodicity and longer run persistence issues for volatility related measures. Another significant early contribution is a largely unnoticed working paper by Dybvig (1991) who explores interest rate volatility through the cumulative sum of squared daily yield changes for the three-month Treasury bill and explicitly refers to it as an empirical version of the quadratic variation process used in analysis of semimartingales. More recently, Zhou (1996) provides an initial study of RV style estimators. He notes that the linkage between sampling frequency and autocorrelation in the high-frequency data series may be induced by sampling noise and he proposes a method to correct for this bias. Andersen and Bollerslev (1997, 1998b) document the simultaneous impact of intraday volatility patterns, the volatility shocks due to macroeconomic news announcements, and the long-run dependence in realized volatility series through an analysis of the cumulative absolute and squared five-minute returns for the Deutsche Mark-Dollar exchange rate. The pronounced intraday features motivate the focus on (multiples of) one trading as the basic aggregation unit for realized volatility measures since this approach largely annihilates repetitive high frequency fluctuations and brings the systematic medium and low frequency volatility variation into focus. Comte and Renault (1998) point to the potential association between RV measures and instantaneous volatility. Finally, early empirical analyses of daily realized volatility measures are provided in, e.g., Andersen et al. (2000b) and Barndorff-Nielsen and Shephard (2001).

7.2 Volatility forecasting

As noted in Section 3, RV is the natural benchmark against which to gauge volatility forecasts. Andersen and Bollerslev (1998a) stress this point which is further developed by Andersen et al. (1999, 2003, 2004) and Patton (2007) through different analytic means.

Several studies pursue alternative approaches in order to improve predictive performance. Ghysels et al. (2006) consider Mixed Data Sampling (MIDAS) regressions that use a combination of volatility measures estimated at different frequencies and horizons. Related, Engle and Gallo (2006) exploit the information in different volatility measures, modelled with a multivariate extension of the multiplicative error model suggested by Engle (2002), to predict multi-step volatility. A rapidly growing literature studies jump detection (e.g., Aït-Sahalia and Jacod (2006), Andersen et al. (2006b, 2007), Fleming et al. (2006), Huang and Tauchen (2005), Jiang and Oomen (2005), Lee and Mykland (2007), Tauchen and Zhou (2006), and Zhang (2007)). Andersen et al. (2007) show that separating the jump and diffusive components in QV estimates enhances the model forecasting performance. Related, Liu and Maheu (2005) and Forsberg and Ghysels (2007) show that realized power variation, which is more robust to the presence of jumps than RV, can improve volatility forecasts.

Other researchers have been investigating the role of microstructure noise on forecasting performance (e.g., Aït-Sahalia and Mancini (2006), Andersen et al. (2005, 2006), and Ghysels and Sinko (2006)) and the issue of how to use noisy overnight return information to enhance volatility forecasts (e.g., Hansen and Lunde (2005) and Fleming et al. (2003)).

A critical feature of volatility is the degree of its temporal dependence. Correlogram plots for the (logarithmic) RV series show a distinct hyperbolic decay that is described well by a fractionally-integrated process. Andersen and Bollerslev (1997) document this feature using the RV series for the Deutsche Mark-Dollar exchange rate. Subsequent studies have documented similar properties across financial markets for the RV on equities (e.g., Andersen et al. (2001), Areal and Taylor (2002), Deo et al. (2006), Martens (2002)), currencies (e.g., Andersen and Bollerslev (1998b), Andersen et al. (2001, 2003), and Zumbach (2004)), and bond yields (e.g., Andersen and Benzoni (2006)). This literature concurs on the value of the fractional integration coefficient, which is estimated in the 0.30–0.48 range, i.e., the stationarity condition is satisfied. Accounting for long memory in volatility can prove useful in forecasting applications (e.g., Deo et al. (2006)). A particularly convenient approach to accommodate the persistent behavior of the RV series is to use a component-based regression to forecast the k-step-ahead quadratic variation (e.g., Andersen et al. (2007), Barndorff-Nielsen and Shephard (2001), and Corsi (2003)):

$$RV(t+k,k) = \beta_0 + \beta_D RV(t,1) + \beta_W RV(t,5) + \beta_M RV(t,21) + \varepsilon(t+k)\,. \quad (25)$$

Simple OLS estimation yields consistent estimates for the coefficients in the regression (25), which can be used to forecast volatility out of sample.

7.3 The distributional implications of the no-arbitrage condition

Equation (19) implies that, approximately, the daily return $r(t)$ follows a Gaussian mixture directed by the IV process. This is reminiscent of the mixture-of-distributions hypothesis analyzed by, e.g., Clark (1973) and Tauchen and Pitts (1983). However, in the case of equation (19) the mixing variable is directly measurable by the RV estimator which facilitates testing the distributional restrictions implied by the no-arbitrage condition embedded in the return dynamics (10). Andersen et al. (2000b) and Thomakos and Wang (2003) find that returns standardized by RV are closer to normal than the standardized residuals from parametric SV models estimated at the daily frequency. Any remaining deviation from normality may be due to a bias in RV stemming from microstructure noise or model misspecification. In particular, when returns jump as in equation (21), or if volatility and return innovations correlate, condition (19) no longer holds. Peters and de Vilder (2006) deal with the volatility-return dependence by sampling returns in 'financial time,' i.e., they identify calendar periods that correspond to equal increments to IV, while Andersen et al. (2007) extend their approach for the presence of jumps. Andersen et al. (2006b) apply these insights, in combination with alternative jump-identification techniques, to different data sets and find evidence consistent with the mixing condition. Along the way they document the importance of jumps and the asymmetric return-volatility relation. Similar issues are also studied in Fleming et al. (2006) and Maheu and McCurdy (2002).

7.4 Multivariate quadratic variation measures

A growing number of studies uses multivariate versions of realized volatility estimators, i.e., realized covariance matrix measures, in portfolio choice (e.g., Bandi et al. (2007) and Fleming et al. (2003)) and risk measurement problems (e.g., Andersen et al. (2001, 2005) and Bollerslev and Zhang (2003)). Multivariate applications, however, are complicated by delays in the security price reactions to price changes in related assets as well as by non-synchronous trading effects. Sheppard (2006) discusses this problem but how to best deal with it remains largely an open issue. Similar to Scholes and Williams (1977), some researchers include temporal cross-correlation terms estimated with lead and lag return data in covariance measures (e.g., Hayashi and Yoshida

(2005, 2006) and Griffin and Oomen (2006)). Other studies explicitly trade off efficiency and noise-induced bias in realized covariance estimates (e.g., Bandi and Russell (2005) and Zhang (2006)), while Bauer and Vorkink (2006) propose a latent-factor model of the realized covariance matrix.

7.5 Realized volatility, model specification and estimation

RV gives empirical content to the latent variance variable and is therefore useful for specification testing of the restrictions imposed on volatility by parametric models previously estimated with low-frequency data. For instance, Andersen and Benzoni (2006) examine the linkage between the quadratic variation and level of bond yields embedded in some affine term structure models and reject the condition that volatility is spanned by bond yields in the U.S. Treasury market. Christoffersen et al. (2006) reject the Heston (1993) model implication that the standard deviation dynamics are conditionally Gaussian by examining the distribution of the changes in the square-root RV measure for S&P 500 returns.

Further, RV measures facilitate direct estimation of parametric models. Barndorff-Nielsen and Shephard (2002) decompose RV into actual volatility and realized volatility error. They consider a state-space representation for this decomposition and apply the Kalman filter to estimate different flavors of the SV model. Bollerslev and Zhou (2002) and Garcia et al. (2001) build on the results of Meddahi (2002b) to obtain efficient moment conditions which they use in the estimation of continuous-time stochastic volatility processes. Todorov (2006b) extends the analysis for the presence of jumps.

8 Possible Directions for Future Research

In recent years the market for derivative securities offering a pure play on volatility has grown rapidly in size and complexity. Well-known examples are the over-the-counter markets for variance swaps, which at maturity pay the difference between realized variance and a fixed strike price, and volatility swaps with payoffs linked to the square root of realized variance. These financial innovations have opened the way for new research on the pricing and hedging of these contracts. For instance, while variance swaps admit a simple replication strategy through static positions in call and put options combined with dynamic trading in the underlying asset (e.g., Britten-Jones and Neuberger (2000) and Carr and Madan (1998)), it is still an open issue to determine the appropriate replication strategy for volatility swaps and other derivatives that are non-linear functions of realized variance (e.g., call and

put options). Carr and Lee (2007) make an interesting contribution in this direction.

Realized volatility is also a useful source of information to learn more about the volatility risk premium. Recent contributions have explored the issue by combining RV measures with model-free option-implied volatility gauges like the VIX (e.g., Bollerslev et al. (2004), Carr and Wu (2007), and Todorov (2006b)). Other studies are examining the linkage between volatility risk and equity premia (Bollerslev and Zhou (2007)), bond premia (Wright and Zhou (2007)), credit spreads (Tauchen and Zhou (2007) and Zhang et al. (2005)), and hedge-fund performance (Bondarenko (2004)). In addition, new research is studying the pricing of volatility risk in individual stock options (e.g., Bakshi and Kapadia (2003), Carr and Wu (2007), Driessen et al. (2006), and Duarte and Jones (2007)) and in the cross section of stock returns (e.g., Ang et al. (2006, 2008), Bandi et al. (2008), and Guo et al. (2007)).

Finally, more work is needed to better understand the linkage between asset return volatility and fluctuations in underlying fundamentals. Several studies have proposed general equilibrium models that generate low-frequency conditional heteroskedasticity (e.g., Bansal and Yaron (2004), Campbell and Cochrane (1999), McQueen and Vorkink (2004), and Tauchen (2005)). Related, Engle and Rangel (2006) and Engle et al. (2006) link macroeconomic variables and long-run volatility movements. An attempt to link medium and higher frequency realized volatility fluctuations in the bond market to both business cycle variation and macroeconomic news releases is initiated in Andersen and Benzoni (2007), but clearly much more work on this front is warranted.

References

Aït-Sahalia, Y. and Jacod, J. (2006): Testing for jumps in a discretely observed process. *Working Paper, Princeton University and Université de Paris-6.*

Aït-Sahalia, Y. and Mancini, L. (2006): Out of sample forecasts of quadratic variation. *Working Paper, Princeton University and University of Zürich.*

Aït-Sahalia, Y., Mykland, P.A. and Zhang, L. (2006): Ultra high frequency volatility estimation with dependent microstructure noise. *Working Paper, Princeton University.*

Aït-Sahalia, Y., Mykland, P.A. and Zhang, L. (2005): How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* **18**, 351–416.

Alizadeh, S., Brandt, M.W. and Diebold, F.X. (2002): Range-based estimation of stochastic volatility models. *Journal of Finance* **57**, 1047–1091.

Andersen, T.G. and Benzoni, L. (2006): Do bonds span volatility risk in the U.S. Treasury market? A specification test for affine term structure models. *Working Paper, KGSM and Federal Reserve Bank of Chicago.*

Andersen, T.G. and Benzoni, L. (2007): The determinants of volatility in the U.S. Treasury market. *Working Paper, KGSM and Federal Reserve Bank of Chicago.*

Andersen, T.G. and Bollerslev, T. (1997): Heterogeneous information arrivals and return volatility dynamics: uncovering the long-run in high frequency returns. *Journal of Finance* **52**, 975–1005.

Andersen, T.G. and Bollerslev, T. (1998a): Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**, 885–905.

Andersen, T.G. and Bollerslev, T. (1998b): Deutsche Mark-Dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* **53**, 219–265.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2004): Parametric and nonparametric volatility measurement. *In: Hansen, L.P. and Aït-Sahalia, Y. (Eds): Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2007): Roughing it up: including jump components in measuring, modeling and forecasting asset return volatility. *Review of Economics and Statistics* forthcoming.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. (2001): The distribution of realized stock return volatility. *Journal of Financial Economics* **61**, 43–76.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2000a): Great realizations. *Risk* **13**, 105–108.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2000b): Exchange rate returns standardized by realized volatility are (nearly) Gaussian. *Multinational Finance Journal* **4**, 159–179.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2001): The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* **96**, 42–55.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003): Modeling and forecasting realized volatility. *Econometrica* **71**, 579–625.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Wu, G. (2005): A framework for exploring the macroeconomic determinants of systematic risk. *American Economic Review* **95**, 398–404.

Andersen, T.G., Bollerslev, T. and Dobrev, D. (2007): No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: theory and testable distributional implications. *Journal of Econometrics* **138**, 125–180.

Andersen, T.G., Bollerslev, T., Frederiksen, P.H. and Nielsen, M.Ø. (2006a): Comment on realized variance and market microstructure noise. *Journal of Business & Economic Statistics* **24**, 173–179.

Andersen, T.G., Bollerslev, T., Frederiksen, P.H. and Nielsen, M.Ø. (2006b): Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. *Working Paper, KGSM*

Andersen, T.G., Bollerslev, T. and Lange, S. (1999): Forecasting financial market volatility: sample frequency vis-á-vis forecast horizon. *Journal of Empirical Finance* **6**, 457–477.

Andersen, T.G., Bollerslev, T. and Meddahi, N. (2004): Analytic evaluation of volatility forecasts. *International Economic Review* **45**, 1079–1110.

Andersen, T.G., Bollerslev, T. and Meddahi, N. (2005): Correcting the errors: volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica* **73**, 279–296.

Andersen, T.G., Bollerslev, T. and Meddahi, N. (2006): Market microstructure noise and realized volatility forecasting. *Working Paper, KGSM.*

Ang, A., Hodrick, R.J., Xing, Y. and Zhang, X. (2006): The cross-section of volatility and expected returns. *Journal of Finance* **51**, 259–299.

Ang, A., Hodrick, R.J., Xing, Y. and Zhang, X. (2008): High idiosyncratic volatility and low returns: international and further U.S. evidence. *Journal of Financial Economics*, forthcoming.

Areal, N.M.P.C. and Taylor, S.J. (2002): The realized volatility of FTSE-100 futures prices. *Journal of Futures Markets* **22**, 627–648.

Back, K. (1991): Asset prices for general processes. *Journal of Mathematical Economics* **20**, 371–395.

Bakshi, G. and Kapadia, N. (2003): Delta-hedged gains and the negative market volatility risk premium. *Review of Financial Studies* **16**, 527–566.

Bandi, F., Moise, C.E. and Russell, J.R. (2008): Market volatility, market frictions, and the cross section of stock returns. *Working Paper, University of Chicago and Case Western Reserve University.*

Bandi, F. and Russell, J.R. (2005): Realized covariation, realized beta, and microstructure noise. *Working Paper, University of Chicago.*

Bandi, F. and Russell, J.R. (2007): Microstructure noise, realized volatility, and optimal sampling. *Review of Economic Studies*, forthcoming.

Bandi, F., Russell, J.R. and Zhu, J. (2007): Using high-frequency data in dynamic portfolio choice. *Econometric Reviews*, forthcoming.

Bansal, R. and Yaron, A. (2004): Risks for the long run: a potential resolution of asset pricing puzzles. *Journal of Finance* **59**, 1481–1509.

Barndorff-Nielsen, O.E. and Shephard, N. (2001): Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics. *Journal of the Royal Statistical Society Series B* **63**, 167–241.

Barndorff-Nielsen, O.E. and Shephard, N. (2002): Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B* **64**, 253–80.

Barndorff-Nielsen, O.E. and Shephard, N. (2004a): Econometric analysis of realised covariation: high-frequency covariance, regression and correlation in financial econometrics. *Econometrica* **72**, 885–925.

Barndorff-Nielsen, O.E. and Shephard, N. (2004b): Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* **2**, 1–37.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N. (2006a): Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. *Working Paper, Aarhus, Stanford, and Oxford Universities.*

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N. (2006b): Subsampling realized kernels. *Working Paper, Aarhus, Stanford, and Oxford Universities.*

Bauer, G.H. and Vorkink, K. (2006): Multivariate realized stock market volatility. *Working Paper, Bank of Canada and Brigham Young University.*

Bollerslev, T., Gibson, M. and Zhou, H. (2004): Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities. *Working Paper, Duke University and Federal Reserve Board of Governors.*

Bollerslev, T. and Zhang, B.Y.B. (2003): Measuring and modeling systematic risk in factor pricing models using high-frequency data. *Journal of Empirical Finance* **10**, 533–558.

Bollerslev, T. and Zhou, H. (2002): Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics* **109**, 33–65.

Bollerslev, T. and Zhou, H. (2007): Expected stock returns and variance risk premia. *Working Paper, Duke University and Federal Reserve Board of Governors.*

Bondarenko, O. (2004): Market price of variance risk and performance of hedge funds. *Working Paper, UIC.*

Brandt, M.W. and Diebold, F.X. (2006): A no-arbitrage approach to range-based estimation of return covariances and correlations. *Journal of Business* **79**, 61–73.

Brandt, M.W. and Jones, C.S. (2006): Volatility forecasting with range-based EGARCH models. *Journal of Business & Economic Statistics* **24**, 470–486.

Britten-Jones, M. and Neuberger, A. (2000): Option prices, implied price processes, and stochastic volatility. *Journal of Finance* **55**, 839–866.

Campbell, J.Y. and Cochrane, J.H. (1999): By force of habit: a consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* **107**, 205–251.

Carr, P. and Lee, R. (2007): Robust replication of volatility derivatives. *Working Paper, NYU and UofC.*

Carr, P. and Madan, D. (1998): Towards a theory of volatility trading. *In: Jarrow, R. (Ed): Volatility.* Risk Publications.

Carr, P. and Wu, L. (2007): Variance risk premia. *Review of Financial Studies*, forthcoming.
Christensen, K. and Podolskij, M. (2006): Realised range-based estimation of integrated variance. *Journal of Econometrics*, forthcoming.
Christoffersen, P.F., Jacobs, K. and Mimouni, K. (2006): Models for S&P 500 dynamics: evidence from realized volatility, daily returns, and option prices. *Working Paper, Mcgill University.*
Clark, P.K. (1973): A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* **41**, 135–155.
Comte, F. and Renault, E. (1998): Long memory in continuous-time stochastic volatility models. *Mathematical Finance* **8**, 291–323.
Corsi, F. (2003): A simple long memory model of realized volatility. *Working paper, University of Southern Switzerland.*
Dacorogna, M.M., Gencay, R., Müller, U., Olsen, R.B. and Pictet, O.V. (2001): *An introduction to high-frequency finance.* Academic Press, San Diego.
Deo, R., Hurvich, C. and Lu, Y. (2006): Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. *Journal of Econometrics* **131**, 29–58.
Dobrev, D. (2007): Capturing volatility from large price moves: generalized range theory and applications. *Working Paper, Federal Reserve Board of Governors.*
Driessen, J., Maenhout, P. and Vilkov, G. (2006): Option-implied correlations and the price of correlation risk. *Working Paper, University of Amsterdam and INSEAD.*
Duarte, J. and Jones, C.S. (2007): The price of market volatility risk. *Working Paper, University of Washington and USC.*
Dybvig, P.H. (1991): Exploration of interest rate data. *Working Paper, Olin School of Business, Washington University in St. Louis.*
Engle, R.F. (2002): New frontiers for ARCH models. *Journal of Applied Econometrics* **17**, 425–446.
Engle, R.F. and Gallo, G.M. (2006): A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* **131**, 3–27.
Engle, R.F., Ghysels, E. and Sohn, B. (2006): On the economic sources of stock market volatility. *Working Paper, NYU and UNC.*
Engle, R.F. and Rangel, J.G. (2006): The spline-GARCH model for low frequency volatility and its global macroeconomic causes. *Working Paper, NYU.*
Fleming, J., Kirby, C. and Ostdiek, B. (2003): The economic value of volatility timing using realized volatility. *Journal of Financial Economics* **67**, 473–509.
Fleming, J. and Paye, B.S. (2006): High-frequency returns, jumps and the mixture-of-normals hypothesis. *Working Paper, Rice University.*
Forsberg, L. and Ghysels, E. (2007): Why do absolute returns predict volatility so well? *Journal of Financial Econometrics* **5**, 31–67.
Foster, D.P. and Nelson, D.B. (1996): Continuous record asymptotics for rolling sample variance estimators *Econometrica* **64**, 139–174.
Gallant, A.R., Hsu, C. and Tauchen, G.E. (1999): Using daily range data to calibrate volatility diffusions and extract the forward integrated variance. *Review of Economics and Statistics* **81**, 617–631.
Garcia, R., Lewis, M.A., Pastorello, S. and Renault, E. (2001): Estimation of objective and risk-neutral distributions based on moments of integrated volatility *Working Paper, Université de Montréal, Banque Nationale du Canada, Universitàà di Bologna, UNC.*
Garman, M.B. and Klass, M.J. (1980): On the estimation of price volatility from historical data. *Journal of Business* **53**, 67–78.
Ghysels, E., Santa-Clara, P. and Valkanov, R. (2006): Predicting volatility: how to get the most out of returns data sampled at different frequencies. *Journal of Econometrics* **131**, 59–95.
Ghysels, E. and Sinko, A. (2006): Volatility forecasting and microstructure noise. *Working Paper, UNC.*

Griffin, J.E. and Oomen, R.C.A. (2006): Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Working Paper, University of Warwick*.

Guo, H., Neely, C.J. and Higbee, J. (2007): Foreign exchange volatility is priced in equities. *Financial Management* forthcoming.

Hansen, P.R. and Lunde, A. (2005): A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics* **3**, 525–554.

Hansen, P.R. and Lunde, A. (2006): Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* **24**, 127–161.

Hayashi, T. and Yoshida, N. (2005): On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* **11**, 359–379.

Hayashi, T. and Yoshida, N. (2006): Estimating correlations with nonsynchronous observations in continuous diffusion models. *Working Paper, Columbia University and University of Tokyo*.

Heston, S.L. (1993): A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* **6**, 327–343.

Hsieh, D.A. (1991): Chaos and nonlinear dynamics: application to financial markets. *Journal of Finance* **46**, 1839–1877.

Huang, X. and Tauchen, G. (2005): The relative contribution of jumps to total price variation. *Journal of Financial Econometrics* **3**, 456–499.

Jacod, J. (1992): Limit of random measures associated with the increments of a Brownian semimartingale. *Working Paper, Publ. Lab. Probabilit N. 120, Paris 6*.

Jiang, G.J. and Oomen, R.C.A. (2005): A new test for jumps in asset prices. *Working Paper, University of Arizona and Warwick Business School*.

Lee, S. and Mykland, P.A. (2007): Jumps in financial markets: a new nonparametric test and jump clustering. *Review of Financial Studies* forthcoming.

Liu, C. and Maheu, J.M. (2005): Modeling and forecasting realized volatility: the role of power variation. *Working Paper, University of Toronto*.

Maheu, J.M. and McCurdy, T.H. (2002): Nonlinear features of realized FX volatility. *Review of Economics and Statistics* **84**, 668–681.

Martens, M. (2002): Measuring and forecasting S&P 500 index-futures volatility using high-frequency data. *Journal of Futures Market* **22**, 497–518.

McQueen, G. and Vorkink, K. (2004): Whence GARCH? A preference-based explanation for conditional volatility. *Review of Financial Studies* **17**, 915–949.

Meddahi, N. (2002a): A theoretical comparison between integrated and realized volatilities. *Journal of Applied Econometrics* **17**, 479–508.

Meddahi, N. (2002b): Moments of continuous time stochastic volatility models. *Working Paper, Imperial College*.

Merton, R.C. (1980): On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics* **8**, 323–361.

Mykland, P.A. (2006): A Gaussian calculus for inference from high frequency data. *Working Paper, University of Chicago*.

Parkinson, M. (1980): The extreme value method for estimating the variance of the rate of return. *Journal of Business* **53**, 61–65.

Patton, A.J. (2007): Volatility forecast comparison using imperfect volatility proxies. *Working Paper, Oxford University*.

Peters, R.T. and de Vilder, R.G. (2006): Testing the continuous semimartingale hypothesis for the S&P 500. *Journal of Business & Economic Statistics* **24**, 444–454.

Protter, P. (1990): *Stochastic Integration and Differential Equations. A New Approach*. Springer, Berlin Heidelberg.

Scholes, M. and Williams, J. (1977): Estimating betas from nonsynchronous data. *Journal of Financial Economics* **5**, 309–327.

Schwert, G.W. (1990): Stock volatility and the crash of '87. *Review of Financial Studies* **3**, 77–102.

Sheppard, K. (2006): Realized covariance and scrambling. *Working Paper, University of Oxford.*

Tauchen, G.E. and Pitts, M. (1983): The price variability-volume relationship on speculative markets. *Econometrica* **51**, 485–505.

Tauchen, G.E. (2005): Stochastic volatility in general equilibrium. *Working Paper, Duke University.*

Tauchen, G.E. and Zhou, H. (2006): Identifying realized jumps on financial markets. *Working Paper, Duke University and Federal Reserve Board of Governors.*

Tauchen, G.E. and Zhou, H. (2007): Realized jumps on financial markets and predicting credit spreads. *Working Paper, Duke University and Federal Reserve Board of Governors.*

Thomakos, D.D. and Wang, T. (2003): Realized volatility in the futures markets. *Journal of Empirical Finance* **10**, 321–353.

Todorov, V. (2006a): Variance risk premium dynamics. *Working Paper, KGSM.*

Todorov, V. (2006b): Estimation of continuous-time stochastic volatility models with jumps using high-frequency data. *Working Paper, KGSM.*

Wright, J. and Zhou, H. (2007): Bond risk premia and realized jump volatility. *Working Paper, Federal Reserve Board of Governors.*

Yang, D. and Zhang, Q. (2000): Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business* **73**, 477–491.

Zhang, L. (2006): Estimating covariation: Epps effect, microstructure noise. *Working Paper, UIC.*

Zhang, L., Mykland, P.A. and Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* **100**, 1394–1411.

Zhang, L. (2007): What you don't know cannot hurt you: on the detection of small jumps. *Working Paper, UIC.*

Zhang, B.Y., Zhou, H. and Zhu, H. (2005): Explaining credit default swap spreads with the equity volatility and jump risks of individual firms. *Working Paper, Fitch Ratings, Federal Reserve Board of Governors, BIS.*

Zhou, B. (1996): High-frequency data and volatility in foreign exchange rates. *Journal of Business and Economic Statistics* **14**, 45–52.

Zumbach, G. (2004): Volatility processes and volatility forecasts with long memory. *Quantitative Finance* **4**, 70–86.

Estimating Volatility in the Presence of Market Microstructure Noise: A Review of the Theory and Practical Considerations

Yacine Aït-Sahalia and Per A. Mykland*

Abstract This chapter reviews our recent work on disentangling high frequency volatility estimators from market microstructure noise, based on maximum-likelihood in the parametric case and two (or more) scales realized volatility (TSRV) in the nonparametric case. We discuss the basic theory, its extensions and the practical implementation of the estimators.

1 Introduction

In this chapter, we review our recent work on disentangling volatility estimators from the market microstructure noise that permeates them at high frequency; the other chapters describe alternative approaches.

We will describe the two complementary estimation strategies that we have developed to decompose asset returns' total variance into one due to the fundamental price and one due to market microstructure noise. The starting point of this analysis is a representation of the observed transaction log price, Y, as the sum of an unobservable efficient price, X, and some noise component due to the imperfections of the trading process, ε :

Yacine Aït-Sahalia
Princeton University and NBER, Bendheim Center for Finance, Princeton University, 26 Prospect Avenue, Princeton, NJ 08540-5296, U.S.A., e-mail: yacine@princeton.edu

Per A. Mykland
Department of Statistics, The University of Chicago, 5734 University Avenue, Chicago, Illinois 60637, U.S.A., e-mail: mykland@pascal.uchicago.edu

* Financial support from the NSF under grants DMS-0532370 (Aït-Sahalia) and DMS 06-04758, and SES 06-31605 (Mykland) is gratefully acknowledged. Computer code in Matlab to implement the estimators described in this chapter is available from the authors upon request.

DOI: 10.1007/978-3-540-71297-8_25,

$$Y_t = X_t + \varepsilon_t. \tag{1}$$

In financial econometrics, one is often interested in estimating the volatility of the efficient log-price process

$$dX_t = \mu_t dt + \sigma_t dW_t \tag{2}$$

using discretely sampled data on the transaction price process at times 0, Δ,..., $n\Delta = T$. In this discussion, we shall assume that there are no jumps.[2]

We will use below two classes of consistent estimators designed for the two situations, one where $\sigma_t \equiv \sigma$ is constant, a fixed parameter to be estimated, and one where σ_t is nonparametric (i.e., an unrestricted stochastic process), in which case we seek to estimate the quadratic variation of the process X, $\int_0^T \sigma_t^2 dt$, over a fixed interval of time T, say one day. Interestingly, the estimator we propose in the parametric case provides a consistent estimator of the quadratic variation of X when the volatility is stochastic.

In both cases, we are also interested in estimating consistently $a^2 = E[\varepsilon^2]$. In some circumstances, the standard deviation of the noise term can be taken as a measure of the liquidity of the market, or the quality of the trade execution in a given exchange or market structure.

For the parametric case, we will focus on the maximum-likelihood estimator developed in Aït-Sahalia et al. (2005a). For the nonparametric case, we will discuss the estimators called Two Scales Realized Volatility (TSRV), which is the first estimator shown to be consistent for $\langle X, X \rangle_T$ (see Zhang et al. (2005b)), and its extension to Multiple Scales Realized Volatility (MSRV) in Zhang (2006).

While we focus on these estimators, others are available. In the constant σ case, French and Roll (1986) proposed to adjust variance estimates to control for the autocorrelation induced by the noise and Harris (1990) studied the resulting estimators. Zhou (1996) proposed a bias correcting approach based on the first order autocovariances, which is unbiased but inconsistent. The behavior of this estimator has been studied by Zumbach et al. (2002). Hansen and Lunde (2006) proposed extensions of the Zhou estimator. A further generalization is provided by Barndorff-Nielsen et al. (2006).

We start in Section 2 with a brief review of the basic theory. Then in Section 3 we discuss many possible refinements to basic theory, including items such as serial correlation in the noise, possible correlation with the underlying price process, etc. In Section 4, we discuss the practical aspects related to the implementation of the methods to actual financial data. Section 5 concludes.

[2] For discussions of jumps in this context, see, for example, Mancini (2001), Aït-Sahalia (2002), Aït-Sahalia (2004), Aït-Sahalia and Jacod (2004), Aït-Sahalia and Jacod (2007), Barndorff-Nielsen and Shephard (2004), Lee and Mykland (2006).

2 Estimators

The basic principle that underlies our analysis is that all the data, despite the fact that it is noisy, should be used when estimating the volatility of the process. This is in direct contrast with the practice in the empirical literature up to now, whereby an arbitrary time interval (say, every few minutes) is selected that is substantially longer than that of the original observations (every few seconds). In order to avoid the high frequency, presumably noisier, observations, this practice entails discarding a substantial portion of the sample.

Our interest in this area started from the realization that one should be able to do better by exploiting the full sample. The estimators we discuss below make full use of the data.

2.1 The parametric volatility case

Consider first the case where σ^2 is constant. The estimation problem is then a case of parametric inference. We here discuss inference when the time horizon $T \to \infty$. The fixed T case is discussed at the end of this section, and in Section 2.2. There is, of course, a large literature on inference for other parametric models as $T \to \infty$, but a review of this is beyond the scope of this article.

If no market microstructure noise were present, i.e., $\varepsilon \equiv 0$, the log-returns $R_i = Y_{\tau_i} - Y_{\tau_{i-1}}$ would be iid $N(0, \sigma^2 \Delta)$. The MLE for σ^2 then coincides with the realized volatility of the process,

$$\widehat{\sigma}^2 = \frac{1}{T} \sum\nolimits_{i=1}^{n} R_i^2. \tag{3}$$

Furthermore, $T^{1/2} \left(\widehat{\sigma}^2 - \sigma^2 \right) \underset{n \longrightarrow \infty}{\longrightarrow} N(0, 2\sigma^4 \Delta)$ and thus selecting Δ as small as possible is optimal for the purpose of estimating σ^2.

When the observations are noisy, with the $\varepsilon' s$ being iid noise with mean 0 and variance a^2, the true structure of the observed log-returns R_i is given by an MA(1) process since

$$R_i = \sigma \left(W_{\tau_i} - W_{\tau_{i-1}} \right) + \varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}} \equiv u_i + \eta u_{i-1} \tag{4}$$

where the $u' s$ are mean zero and variance γ^2 with

$$\gamma^2 (1 + \eta^2) = \mathrm{Var}[R_i] = \sigma^2 \Delta + 2a^2 \tag{5}$$

$$\gamma^2 \eta = \mathrm{Cov}(R_i, R_{i-1}) = -a^2. \tag{6}$$

If we assume for a moment that $\varepsilon \sim N(0, a^2)$ (an assumption we will relax below), then the $u' s$ are iid Gaussian and the likelihood function for the vector

R of observed log-returns, as a function of the transformed parameters (γ^2, η), is given by

$$l(\eta, \gamma^2) = -\ln \det(V)/2 - n\ln(2\pi\gamma^2)/2 - (2\gamma^2)^{-1} R' \Omega^{-1} R \tag{7}$$

where

$$\Omega = [\omega_{ij}] = \begin{pmatrix} 1+\eta^2 & \eta & \cdots & 0 \\ \eta & 1+\eta^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \eta \\ 0 & \cdots & \eta & 1+\eta^2 \end{pmatrix}.$$

From the perspective of practical implementation, this estimator is nothing else than the MLE estimator of an MA(1) process with Gaussian errors: any existing computer routines for the MA(1) situation can therefore be applied (see e.g., Section 5.4 in Hamilton (1995)). In particular, the log-likelihood function can be expressed in a computationally efficient form by triangularizing the matrix Ω, avoiding the brute-force computation of Ω^{-1} and yielding the equivalent expression:

$$l(\eta, \gamma^2) = -\frac{1}{2}\sum_{i=1}^{N} \ln(2\pi d_i) - \frac{1}{2}\sum_{i=1}^{N} \frac{\widetilde{Y}_i^2}{d_i}, \tag{8}$$

where

$$d_i = \gamma^2 \frac{1+\eta^2+...+\eta^{2i}}{1+\eta^2+...+\eta^{2(i-1)}}$$

and the $\widetilde{Y}_i's$ are obtained recursively as $\widetilde{Y}_1 = Y_1$ and for $i = 2, ..., N$:

$$\widetilde{Y}_i = Y_i - \frac{\eta\left(1+\eta^2+...+\eta^{2(i-2)}\right)}{1+\eta^2+...+\eta^{2(i-1)}}\widetilde{Y}_{i-1}.$$

The MLE $(\widehat{\sigma}^2, \widehat{a}^2)$ is consistent and its asymptotic variance is given by

$$\mathrm{AVAR}_{\mathrm{normal}}(\widehat{\sigma}^2, \widehat{a}^2) = \begin{pmatrix} 4\left(\sigma^6\Delta\left(4a^2+\sigma^2\Delta\right)\right)^{1/2} + 2\sigma^4\Delta & -\sigma^2\Delta h \\ \bullet & \frac{\Delta}{2}\left(2a^2+\sigma^2\Delta\right)h \end{pmatrix}$$

with

$$h \equiv 2a^2 + \left(\sigma^2\Delta\left(4a^2+\sigma^2\Delta\right)\right)^{1/2} + \sigma^2\Delta.$$

Since $\mathrm{AVAR}_{\mathrm{normal}}(\widehat{\sigma}^2)$ is increasing in Δ, we are back to the situation where it is optimal to sample as often as possible. Interestingly, the AVAR structure of the estimator remains largely intact if we misspecify the distribution of the microstructure noise. Our likelihood function then becomes a quasi-likelihood in the sense of Wedderburn (1974), see also McCullagh and Nelder (1989).

Specifically, suppose that the $\varepsilon's$ have mean 0 and variance a^2 but are not normally distributed. If the econometrician (mistakenly) assumes that the $\varepsilon's$

are normal, inference is still done with the Gaussian log-likelihood $l(\sigma^2, a^2)$, using the scores $\dot{l}_{\sigma^2}$ and $\dot{l}_{a^2}$ as moment functions. Since the expected values of $\dot{l}_{\sigma^2}$ and $\dot{l}_{a^2}$ only depend on the second order moment structure of the log-returns R, which is unchanged by the absence of normality, the moment functions are unbiased:

$$E_{\text{true}}[\dot{l}_{\sigma^2}] = E_{\text{true}}[\dot{l}_{a^2}] = 0$$

where "true" denotes the true distribution of the $Y's$. Hence the estimator $(\widehat{\sigma}^2, \widehat{a}^2)$ based on these moment functions remains consistent and the effect of misspecification lies in the AVAR. By using the cumulants of the distribution of ε, we express the AVAR in terms of deviations from normality. We obtain that the estimator $(\widehat{\sigma}^2, \widehat{a}^2)$ is consistent and its asymptotic variance is given by

$$\text{AVAR}_{\text{true}}(\widehat{\sigma}^2, \widehat{a}^2) = \text{AVAR}_{\text{normal}}(\widehat{\sigma}^2, \widehat{a}^2) + \text{Cum}_4[\varepsilon] \begin{pmatrix} 0 & 0 \\ 0 & \Delta \end{pmatrix}$$

where $\text{AVAR}_{\text{normal}}(\widehat{\sigma}^2, \widehat{a}^2)$ is the asymptotic variance in the case where the distribution of U is Normal. ε has mean zero, so in terms of its moments

$$\text{Cum}_4[\varepsilon] = E\left[\varepsilon^4\right] - 3\left(E\left[\varepsilon^2\right]\right)^2. \tag{9}$$

In the special case where ε is normally distributed, $\text{Cum}_4[\varepsilon] = 0$.

The i.i.d. assumption for noise may not be realistic. We discuss serially correlated noise below. For a study of that happens if the noise can depend on the latent process, we refer to Li and Mykland (2007).

The presence of a drift does not alter these earlier conclusions, not just because it would be economically irrelevant at the observation frequencies we consider, but also because of the following. Suppose that $X_t = \mu t + \sigma W_t$ then the block of the AVAR matrix corresponding to $(\widehat{\sigma}^2, \widehat{a}^2)$ is the same as if μ were known, in other words, as if $\mu = 0$, which is the case we focused on.

In the T fixed case, inference for fixed σ has been considered by Gloter (2000) and Gloter and Jacod (2000). The estimator converges at the rate of $n^{1/4}$, where n is the number of observations. An important issue is what happens to this MLE estimator if the volatility is in fact not constant, but is stochastic as in the model which we will now turn to. In that case, the likelihood function becomes misspecified. Recent results in Xiu (2008), however, suggest that the MLE estimator, suitably rescaled, is a consistent estimator of the integrated volatility of the process (see Xiu (2008)). This analysis may explain why the MLE seems to perform so well in simulations done under a variety of non-constant volatility models (see Gatheral and Oomen (2007)).

2.2 The nonparametric stochastic volatility case

An alternative model is one where volatility is stochastic, and we now discuss how to separate the fundamental and noise volatilities in this case. When

$$dX_t = \sigma_t dW_t, \tag{10}$$

the object of interest is now the quadratic variation

$$\langle X, X \rangle_T = \int_0^T \sigma_t^2 dt \tag{11}$$

over a fixed time period $[0, T]$. The usual estimator of $\langle X, X \rangle_T$ is the realized volatility (RV)

$$[Y, Y]_T = \sum_{i=1}^{n} (Y_{t_{i+1}} - Y_{t_i})^2. \tag{12}$$

In the absence of noise, $[Y, Y]_T$ consistently estimates $\langle X, X \rangle_T$. The sum converges to the integral, with a known distribution, dating back to Jacod (1994) and Jacod and Protter (1998). As in the constant σ case, selecting Δ as small as possible ($= n$ as large as possible) is optimal.

But ignoring market microstructure noise leads to an even more dangerous situation than when σ is constant and $T \to \infty$. In the high frequency limit, market microstructure noise totally swamps the variance of the price signal at the level of the realized volatility, as was noted by Zhou (1996). More precisely, after suitable scaling, RV based on the observed log-returns is a consistent and asymptotically normal estimator – but of the quantity $2nE[\varepsilon^2]$. This quantity has nothing to do with the object of interest, $\langle X, X \rangle_T$.

This is of course already visible in the special case of constant volatility we just studied. Since the expressions above are exact small-sample ones, they can in particular be specialized to analyze the situation where one samples at increasingly higher frequency ($\Delta \to 0$, say sampled every minute) over a fixed time period (T fixed, say a day). If we continue to assume iid noise, with $n = T/\Delta$, we have

$$E\left[\widehat{\sigma}^2\right] = \frac{2na^2}{T} + o(n) = \frac{2nE\left[\varepsilon^2\right]}{T} + o(n) \tag{13}$$

$$\text{Var}\left[\widehat{\sigma}^2\right] = \frac{2n\left(6a^4 + 2\,\text{Cum}_4\left[\varepsilon\right]\right)}{T^2} + o(n) = \frac{4nE\left[\varepsilon^4\right]}{T^2} + o(n) \tag{14}$$

so $(T/2n)\widehat{\sigma}^2$ becomes an estimator of $E\left[\varepsilon^2\right] = a^2$ whose asymptotic variance is $E\left[\varepsilon^4\right]$. Note in particular that $\widehat{\sigma}^2$ estimates the variance of the noise, which is essentially unrelated to the object of interest σ^2.

In fact, if one uses all the data (say sampled every second),

$$[Y,Y]_T^{(\text{all})} \overset{\mathcal{L}}{\approx} \underbrace{\langle X,X\rangle_T}_{\text{object of interest}} + \underbrace{2nE[\varepsilon^2]}_{\text{bias due to noise}} + \underbrace{\Big[\underbrace{4nE[\varepsilon^4]}_{\text{due to noise}} + \underbrace{\frac{2T}{n}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}\Big]^{1/2} Z_{\text{total}}}_{\text{total variance}}.$$

conditionally on the X process, where Z denotes a standard normal variable. So the bias term due to the noise, which is of order $O(n)$, swamps the true quadratic variation $\langle X,X\rangle_T$, which is of order $O(1)$.

While a formal analysis of this phenomenon originated in Zhang et al. (2005b), it has long been known that sampling as prescribed by $[Y,Y]_T^{(\text{all})}$ is not a good idea. The recommendation in the literature is to sample sparsely at some lower frequency, by using a realized volatility estimator $[Y,Y]_T^{(\text{sparse})}$ constructed by summing squared log-returns at some lower frequency, usually 5, 10, 15, 30 minutes (see e.g., Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002) and Gençay et al. (2002).) Reducing the value of n, from say $23,400$ (1 second sampling) to 78 (5 minute sampling over the same 6.5 hours), has the advantage of reducing the magnitude of the bias term $2nE[\varepsilon^2]$. Yet, one of the most basic lessons of statistics is that discarding data is, in general, not advisable.

Zhang et al. (2005b) propose a solution to this problem which makes use of the full data sample yet delivers consistent estimators of both $\langle X,X\rangle_T$ and a^2. The estimator, Two Scales Realized Volatility (TSRV), is based on subsampling, averaging and bias-correction. By evaluating the quadratic variation at two different frequencies, averaging the results over the entire sampling, and taking a suitable linear combination of the result at the two frequencies, one obtains a consistent and asymptotically unbiased estimator of $\langle X,X\rangle_T$.

TSRV's construction is quite simple: first, partition the original grid of observation times, $G=\{t_0,...,t_n\}$ into subsamples, $G^{(k)}$, $k=1,...,K$ where $n/K\to\infty$ as $n\to\infty$. For example, for $G^{(1)}$ start at the first observation and take an observation every 5 minutes; for $G^{(2)}$, start at the second observation and take an observation every 5 minutes, etc. Then we average the estimators obtained on the subsamples. To the extent that there is a benefit to subsampling, this benefit can now be retained, while the variation of the estimator will be lessened by the averaging. This reduction in the estimator's variability will open the door to the possibility of doing bias correction.

Averaging over the subsamples gives rise to the estimator

$$[Y,Y]_T^{(\text{avg})} = \frac{1}{K}\sum_{k=1}^{K}[Y,Y]_T^{(k)}$$

constructed by averaging the estimators $[Y,Y]_T^{(k)}$ obtained on K grids of average size $\bar{n} = n/K$. The properties of this estimator are given by

$$[Y,Y]_T^{(\text{avg})} \overset{\mathcal{L}}{\approx} \underbrace{\langle X,X\rangle_T}_{\text{object of interest}} + \underbrace{2\bar{n}E[\varepsilon^2]}_{\text{bias due to noise}} + \underbrace{[\underbrace{4\frac{\bar{n}}{K}E[\varepsilon^4]}_{\text{due to noise}} + \underbrace{\frac{4T}{3\bar{n}}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}]^{1/2} Z_{\text{total}}}_{\text{total variance}}.$$

While a better estimator than $[Y,Y]_T^{(\text{all})}$, $[Y,Y]_T^{(\text{avg})}$ remains biased. The bias of $[Y,Y]_T^{(\text{avg})}$ is $2\bar{n}E[\varepsilon^2]$; of course, $\bar{n} < n$, so progress is being made. But one can go one step further. Indeed, $E[\varepsilon^2]$ can be consistently approximated using RV computed with all the observations:

$$\widehat{E[\varepsilon^2]} = \frac{1}{2n}[Y,Y]_T^{(\text{all})} \tag{15}$$

Hence the bias of $[Y,Y]^{(\text{avg})}$ can be consistently estimated by $\bar{n}[Y,Y]_T^{(\text{all})}/n$. TSRV is the bias-adjusted estimator for $\langle X,X\rangle$ constructed as

$$\widehat{\langle X,X\rangle}_T^{(\text{tsrv})} = \underbrace{[Y,Y]_T^{(\text{avg})}}_{\text{slow time scale}} - \frac{\bar{n}}{n}\underbrace{[Y,Y]_T^{(\text{all})}}_{\text{fast time scale}}. \tag{16}$$

If the number of subsamples is optimally selected as

$$K^* = cn^{2/3}, \tag{17}$$

then TSRV has the following distribution:

$$\widehat{\langle X,X\rangle}_T^{(\text{tsrv})} \overset{\mathcal{L}}{\approx} \underbrace{\langle X,X\rangle_T}_{\text{object of interest}} + \frac{1}{n^{1/6}}\underbrace{[\underbrace{\frac{8}{c^2}E[\varepsilon^2]^2}_{\text{due to noise}} + \underbrace{c\frac{4T}{3}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}]^{1/2} Z_{\text{total}}}_{\text{total variance}}. \tag{18}$$

Unlike all the previously considered ones, this estimator is now correctly centered. The optimal choice of the constant c is given by

$$c^* = \left(\frac{T}{12(E\epsilon^2)^2}\int_0^T \sigma_t^4 dt\right)^{-1/3}. \tag{19}$$

Consistent estimators for that quantity are given in Section 6 of Zhang et al. (2005b).

A small sample refinement to $\widehat{\langle X,X\rangle}_T$ can be constructed as follows

$$\widehat{\langle X,X\rangle}_T^{(\text{tsrv,adj})} = \left(1-\frac{\overline{n}}{n}\right)^{-1} \widehat{\langle X,X\rangle}_T^{(\text{tsrv})}. \tag{20}$$

The difference with the estimator (16) is of order $O_p(K^{-1})$, and thus the two estimators have the same asymptotic behaviors to the order that we consider. However, the estimator (20) is unbiased to higher order.

It should be emphasized that the development above works equally well for non-equidistant observations. One can therefore move to sampling in tick time. The only modification is that the integral $\int_0^T \sigma_t^4 dt$ gets replaced by $\int_0^T \sigma_T^4 dH(t)$, where $H(t)$ is a measure if the quadratic variation of sampling times. For details, we refer to Zhang et al. (2005b).

TSRV provides the first consistent and asymptotic (mixed) normal estimator of the quadratic variation $\langle X,X\rangle_T$; as can be seen from (18), it has the rate of convergence $n^{-1/6}$. Finally, we emphasize that the parametric and nonparametric cases above are different not only in the degree of parametric specification, but also in the form of asymptotics. In the former case, $T\to\infty$, while in the latter, T is fixed at, say, one day. The case of parametric inference for fixed T and high frequency data has been discussed by Gloter (2000) and Gloter and Jacod (2000).

3 Refinements

3.1 Multi-scale realized volatility

If one can benefit from combining two scales, how about combining several? Essentially, using two scales assures consistency and (asymptotic) unbiasedness; adding more scales improves efficiency. Zhang (2006) shows that it is possible to generalize TSRV, by averaging not just on two but on multiple time scales. For suitably selected weights, the resulting estimator, MSRV converges to $\langle X,X\rangle_T$ at the somewhat faster rate $n^{-1/4}$. An related generalization of TSRV is also provided by Barndorff-Nielsen et al. (2006).

TSRV corresponds to the special case where one uses a single slow time scale in conjunction with the fast time scale to bias-correct it. As has been shown in Gloter (2000) and Gloter and Jacod (2000), this is the same rate as for parametric inference, and so the rate is the best attainable also in the nonparametric situation.

We refer to Zhang (2006) for the general description of the Multiple Scales Realized Volatility (MSRV), but note in particular the following: (1) The MSRV estimator is $n^{-1/4}$-consistent even with the usual edge (beginning and end of day) effects. This is not typically the case for estimators based on

autocovariances. (2) The MSRV does depend on a weighting function, which can be chosen to optimize efficiency.

We note that for purposes of efficiency, the constant volatility case is useful as a benchmark even if such constancy does not actually apply. Compared to the TSRV, the MSRV is somewhat harder to implement, and the efficiency gain for typical sample sizes is moderate. On the other hand, the MSRV has the advantage of only involving subsampling over points that are $O(n^{1/2})$ observations apart (while for the TSRV, they can be $O(n^{2/3})$ observations apart). We believe that one can use either estimator with a good conscience.

3.2 Non-equally spaced observations

A substantial fraction of the literature is developed for equidistant observations. This may make sense when one takes subsamples every five or fifteen minutes, but is unrealistic in ultra high frequency data. While irregular spacing does not alter the estimators of volatility (such as TSRV or MSRV), it does alter the standard errors of estimators. In the more realistic case, such standard errors have to involve some notion of the quadratic variation of time, as first documented in Zhang (2001) and Mykland and Zhang (2006). When combining several scales, expressions become more difficult, but are still estimable, see, in particular, the developments in Zhang et al. (2005b).

It is a commonly believed misconception that non-equally spaced data can be analyzed by assuming that the i'th sampling time (in a sample of size n) is given by $t_{n,i} = f(i/n)$, where f is an increasing (possibly random) function which does not depend on n. However, this is just a rescaling of time by the function f, and it does not alter the asymptotic variance of any estimator of daily volatility. This mode of analysis, therefore, only captures very mild forms of irregular spacing. For example, in the case of sampling at Poisson times, the sampling points are uniformly distributed, and the function f, had it existed, would have to be linear. However, the asymptotic quadratic variation of time (AQVT, see Zhang (2001), Zhang et al. (2005b), Mykland and Zhang (2006) and Zhang (2006)) for Poisson sampling is double that of sampling at regularly spaced times.

A somewhat mitigating factor in this picture is that estimates of asymptotic variance in some cases automatically take account of the irregularly of sampling, see, for example, Remark 2 (p. 1944) of Mykland and Zhang (2006). The more general case, as for example in Zhang et al. (2005b), is more complicated, and investigations are continuing. Natural questions are whether trading intensity variation will be associated with time variation in the noise distribution at very high frequency, and sampling in tick time could potentially render the noise distribution more homogenous.

The problem, however, is further mitigated by subsampling. It would seem that for many data generating mechanisms for the sampling times, the averaging of spacings implied by subsampling makes the subsampled times less irregular. For example, if one subsamples every K'th observation time when spacings are from a Poisson process, the AQVT of the subsampled times will decrease to that of regular sampling as K gets larger. This is an additional benefit of subsampling.

3.3 Serially-correlated noise

3.3.1 The parametric case

The likelihood function can be modified in the case of serially correlated noise. The form of the variance matrix of the observed log-returns must be altered, replacing $\gamma^2 v_{ij}$ with

$$\begin{aligned} & \mathrm{Cov}(R_i, R_j) \\ = \; & \mathrm{Cov}(\sigma\left(W_{\tau_i} - W_{\tau_{i-1}}\right) + \varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}}, \sigma\left(W_{\tau_j} - W_{\tau_{j-1}}\right) + \varepsilon_{\tau_j} - \varepsilon_{\tau_{j-1}}) \\ = \; & \sigma^2 \Delta \delta_{ij} + \mathrm{Cov}(\varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}}, \varepsilon_{\tau_j} - \varepsilon_{\tau_{j-1}}) \end{aligned}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. A model for the time series dependence of the ε and its potential correlation to the price process would then specify the remaining terms.

3.3.2 The nonparametric case

When the microstructure noise ϵ is iid, log-returns

$$R_i = Y_{\tau_i} - Y_{\tau_{i-1}} = \int_{\tau_{i-1}}^{\tau_i} \sigma_t dW_t + \epsilon_{\tau_i} - \epsilon_{\tau_{i-1}} \tag{21}$$

follow an MA(1) process since the increments $\int_{\tau_{i-1}}^{\tau_i} \sigma_t dW_t$ are uncorrelated, $\epsilon \perp W$ and therefore, in the simple case where σ_t is nonrandom (but possibly time varying),

$$E\left[R_i R_j\right] = \begin{cases} \int_{\tau_{i-1}}^{\tau_i} \sigma_t^2 dt + 2E\left[\epsilon^2\right] & \text{if } j = i \\ -E\left[\epsilon^2\right] & \text{if } j = i+1 \\ 0 & \text{if } j > i+1 \end{cases} \tag{22}$$

Under the simple i.i.d. noise assumption, log-returns are therefore (negatively) autocorrelated at the first order.

An example of a simple model to capture higher order serial dependence in ϵ is

$$\epsilon_{t_i} = U_{t_i} + V_{t_i} \tag{23}$$

where U is iid, V is $AR(1)$ with first order coefficient ρ, $|\rho| < 1$, and $U \perp V$. Under this model, we have

$$E\left[R_i R_j\right] = \begin{cases} \int_{\tau_{i-1}}^{\tau_i} \sigma_t^2 dt + 2E\left[U^2\right] + 2\left(1-\rho\right) E\left[V^2\right] & \text{if } j = i \\ -E\left[U^2\right] - \left(1-\rho\right)^2 E\left[V^2\right] & \text{if } j = i+1 \\ -\rho^{j-i-1}\left(1-\rho\right)^2 E\left[V^2\right] & \text{if } j > i+1 \end{cases} \tag{24}$$

More generally, assume that the noise process ϵ_{t_i} is independent of the X_t process, and that it is (when viewed as a process in index i) stationary and strong mixing with the mixing coefficients decaying exponentially. We also suppose that for some $\kappa > 0$, $E\epsilon^{4+\kappa} < \infty$.

Definitions of mixing concepts can be found e.g., in Hall and Heyde (1980), p. 132. Note that by Theorem A.6 (p. 278) of Hall and Heyde (1980), there is a constant $\rho < 1$ so that, for all i,

$$\left|\mathrm{Cov}(\epsilon_{t_i}, \epsilon_{t_{i+l}})\right| \;\leq\; \rho^l \,\mathrm{Var}(\epsilon) \tag{25}$$

Note that we are modeling the noise process dependence in tick time.

For the moment, we focus on determining the integrated volatility of X for one time period $[0, T]$. This is also known as the continuous quadratic variation $\langle X, X\rangle$ of X. In other words,

$$\langle X, X\rangle_T = \int_0^T \sigma_t^2 dt. \tag{26}$$

Our volatility estimators can be described by considering subsamples of the total set of observations. A realized volatility based on every j'th observation, and starting with observation number r, is given as

$$[Y, Y]_T^{(j,r)} = \sum_{0 \leq j(i-1) \leq n-r-j} \left(Y_{t_{ji+r}} - Y_{t_{j(i-1)+r}}\right)^2.$$

Under most assumptions, this estimator violates the sufficiency principle, whence we define the *average lag j realized volatility* as

$$\begin{aligned} [Y, Y]_T^{(J)} &= \frac{1}{J}\sum_{r=0}^{J-1} [Y, Y]_T^{(J,r)} \\ &= \frac{1}{J}\sum_{i=0}^{n-J} \left(Y_{t_{i+J}} - Y_{t_i}\right)^2. \end{aligned} \tag{27}$$

A generalization of TSRV can be defined for $1 \leq J < K \leq n$ as

$$\widehat{\langle X,X\rangle}_T^{(\text{tsrv})} = \underbrace{[Y,Y]_T^{(K)}}_{\text{slow time scale}} - \frac{\bar{n}_K}{\bar{n}_J}\underbrace{[Y,Y]_T^{(J)}}_{\text{fast time scale}}, \tag{28}$$

thereby combining the two time scales J and K. Here $\bar{n}_K = (n-K+1)/K$ and similarly for $\bar{n}_J$.

We will continue to call this estimator the TSRV estimator, noting that the estimator we proposed in Zhang et al. (2005b) is the special case where $J=1$ and $K \to \infty$ as $n \to \infty$. This more general estimator remains consistent and asymptotically mixed normal under assumption (25), as discussed in Aït-Sahalia et al. (2005b).

Both (16) and (28) are estimators of the same quantity, the quadratic variation, derived under different assumptions on the correlation structure of the noise (iid vs. fairly time series dependence.) One possibility is to approach the question with an eye towards robsutness considerations: in the Hausman test spirit, if the two estimators are close, then it is likely that the iid assumption for the noise is not a bad one in this particular instance.

3.4 Noise correlated with the price signal

3.4.1 The parametric case

The likelihood function can be modified in the case of noise that is both serially correlated, as discussed above, but also correlated with the price process. In those cases, the form of the variance matrix of the observed log-returns must be altered, replacing $\gamma^2 v_{ij}$ with

$$\begin{aligned}
&\text{Cov}(R_i, R_j) \\
&= \text{Cov}(\sigma\left(W_{\tau_i} - W_{\tau_{i-1}}\right) + \varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}}, \sigma\left(W_{\tau_j} - W_{\tau_{j-1}}\right) + \varepsilon_{\tau_j} - \varepsilon_{\tau_{j-1}}) \\
&= \sigma^2 \Delta\delta_{ij} + \text{Cov}(\sigma\left(W_{\tau_i} - W_{\tau_{i-1}}\right), \varepsilon_{\tau_j} - \varepsilon_{\tau_{j-1}}) \\
&\quad + \text{Cov}(\sigma\left(W_{\tau_j} - W_{\tau_{j-1}}\right), \varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}}) + \text{Cov}(\varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}}, \varepsilon_{\tau_j} - \varepsilon_{\tau_{j-1}})
\end{aligned} \tag{29}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

3.4.2 The nonparametric case

In principle, one could build a complex model to relate microstructure noise to the efficient price . In practice, however, separating the signal from the noise is not that simple, despite claims to the contrary in the literature (see e.g., Hansen and Lunde (2006)). There are several difficult issues concerning how to model the noise. First of all, the noise can only be distinguished from the efficient price under fairly careful modelling. In most cases, the

assumption that the noise is stationary, alone, is not enough to make the noise identifiable. For example, one could write down an additive model for the observed (log) price process $\{p_t\}$:

$$p_{t_{i,m}} = p^*_{t_{i,m}} + u_{t_{i,m}},$$

and denote p^*_t and u_t are, respectively, signal and noise. This model, however, does not guarantee that one can disentangle the signal or the volatility of the signal. To see this, suppose that the efficient price can be written as

$$dp^*_t = \mu_t dt + \sigma_t dW_t,$$

where the drift coefficient μ_t and the diffusion coefficient σ_t can be random, and W_t is a standard Brownian motion. If one assumed that u_t is also an Itô process, say,

$$du_t = \nu_t dt + \gamma_t dB_t,$$

then p_t is also an Itô process of the form

$$dp_t = (\mu_t + \nu_t)dt + \omega_t dV_t,$$

where $\omega^2_t = \sigma^2_t + \gamma^2_t + 2\sigma_t\gamma_t d < W, B >_t /dt$ (by the Kunita-Watanabe inequality, see, for example, Protter (2004)).

Unless one imposes additional constraints, it is therefore *not* possible to distinguish signal and noise in this model, and the integrated variance (quadratic variation) of the process should be taken to be $\int_0^T \omega^2_t dt$. One could, of course, require $\mu_t = 0$, as is done in Aït-Sahalia et al. (2005a), but consistent estimation is only possible as $T \to \infty$, and given a parametric model or similar. Note that many papers make the assumption that the drift in the efficient price is zero. This assumption is necessary for their unbiasedness considerations, but for the purposes of asymptotics in a fixed time interval T such as a day, it does not matter. This is for the same reason that a consistent separation of efficient price is not possible in a fixed time interval, so long as the noise is also an Itô process.

The same statement, broadly interpreted (replace integrated volatility of any process X with its quadratic variation), holds true for general semimartingales, see Theorem I.4.47 (p. 52) of Jacod and Shiryaev (2003). One can in some cases extend the concept of quadratic variation to non-semimartingales. As we shall see, even in this case the noise is not separable from the signal except under additional assumptions.

What makes this problem particularly difficult is that a substantial fraction of continuous processes of interest here are Itô processes. And many Itô processes have stationary solutions. One can easily construct a stationary diffusion process with given marginal distribution and exponential autocorrelation function. By superposition, and by taking limits, one can, for example, construct a Gaussian Itô process with mean zero and with any autocovariance

function on the form $\pi(s) = \int_0^\infty e^{-us}\nu(du)$, where ν is any finite measure on $[0, \infty)$.

The only case where one can hope to distinguish between efficient price and noise, is if the noise u_t is not an Itô process. One way for this to occur is if the u_t are independent for different t, and hence the autocovariance function satisfies $\pi(s) = 0$ for $s \neq 0$. It should be emphasized that consistency is not guaranteed even if the noise is not a semimartingale; this kind of model has to be analyzed in each individual case.

3.5 Small sample edgeworth expansions

Our experience suggests that practitioners trading volatility derivatives such as variance swaps very much care about the degree of uncertainty associated with volatility estimators. Their contracts are written with respect to a specific volatility measure, RV in most cases. To the extent that better estimators of IV are available, they suggest trading opportunities.

In some instances, the asymptotic normal approximation to the error of the estimators described above can be improved on using Edgeworth expansions and Cornish-Fisher inversions. In the context of TSRV, this is explained in Zhang et al. (2005a), to which we refer. In the case where there is no microstructure noise, expansions for feasible estimators are discussed by Goncalves and Meddahi (2005). A good account of the general Edgeworth theory can be found in Hall (1992).

3.6 Robustness to departures from the data generating process assumptions

Actual volatility data can depart from the stylized assumptions we have made above in a number of ways: they may exhibit long memory, jumps, etc. We refer to Aït-Sahalia and Mancini (2006) and Gatheral and Oomen (2007) for various simulations that document the robustness of MLE, TSRV and MSRV in these settings.

4 Computational and Practical Implementation Considerations

4.1 Calendar, tick and transaction time sampling

One argument for tick time sampling is based on sampling more when the market is active, as opposed to sampling, say, once every ten seconds. Griffin and Oomen (2006) provide an interesting analysis of the impact of tick vs. transaction sampling. Their results show that the nature of the sampling mechanism can generate fairly distinct autocorrelogram patterns for the resulting log-returns. Now, from a practical perspective, we can view the choice of sampling scheme as one more source of noise, this one attributable to the econometrician who is deciding between different ways to approach the same original transactions or quotes data: should we sample in calendar time? transaction time? tick time? something else altogether? Since the sampling mechanism is not dictated by the data, this argues for working under robust departures from the basic assumptions.

An additional issue is whether to model time dependence in the microstructure in calendar or tick time. As discussed in Section 3.4.2, modeling such dependence in calendar time results in identifiability problems, which is why we have focused on tick time. It remains an interesting question how to effectively do inference with such dependence in calendar time.

4.2 Transactions or quotes

Quotes do not represent an actual price at which a transaction took place. As such they can be subject to caution in terms of interpretation as a price series. But they contain substantially more information regarding the strategic behavior of market makers, for instance. Overall, the midpoint of the bid and offer quotes data at each point in time, weighted by the quoted market depth for each quote, tends to produce a series that is substantially less affected by market microstructure noise than the transactions price and should probably be used at least for comparison purposes whenever possible.

The model (23) for the microstructure noise describes well a situation where the primary source of the noise beyond order one consists of further bid-ask bounces. In such a situation, the fact that a transaction is on the bid or ask side has little predictive power for the next transaction, or at least not enough to predict that two successive transactions are on the same side with very high probability (although Choi et al. (1988) have argued that serial correlation in the transaction type can be a component of the bid-ask spread, and extended the model of Roll (1984) to allow for it).

In trying to assess the source of the higher order dependence in the log-returns, a natural hypothesis is that this is due to the trade reversals: in transactions data and an orderly liquid market, one might expect that in most cases successive transactions of the same sign (buy or sell orders) will not move the price. The next recorded price move is then, more likely than not, going to be caused by a transaction that occurs on the other side of the bid-ask spread, and so we observed these reversals when the data consist of the transactions that lead to a price change. In other words, when looking at second-to-second price moves, volatility will often be too low to change prices by a full tick. Hence, observed price moves will often be dominated by rounding. This point is in fact what motivates the Roll (1984) estimator of the bid-ask spread.

We looked at quotes data, also from the TAQ database. Indeed, an important source for the AR(1) pattern found in transactions returns with negative autocorrelation (the term V in (23)) will be trade reversals. The remaining autocorrelation exhibited in the quotes data can also be captured by model (23), but with a positive autocorrelation in the V term. This can capture effects such as the gradual adjustment of prices in response to a shock such as a large trade. So the pattens of autocorrelations, beyond the salient MA(1) first order term are quite different between transactions and quotes series.

4.3 Selecting the number of subsamples in practice

The asymptotic theory summarized above provides specific rules for optimally selecting the number of subsamples in TSRV, K, and consequently the average number of observations per subsample, $\bar{n}$. These rules are asymptotic by nature and provide some reasonable guidance in small samples. Implementing them to the letter, however, does require the estimation of additional quantities, such as $\int_0^T \sigma_t^4 dt$: see (17) and (19). This requires additional effort, and is not trivial in the presence of noise, but furthermore it may not be that useful in practice.

In practice, we have found that the best approach is usually to start with reasonable values (say if $n = 23,400$ observations or 1 per second, then start at $K = 300$ or $\bar{n} = 78$ corresponding to subsampling every 5 minutes) and work around those values (in the same example from a range of say 1 minute to 10 minutes): those numbers can be adjusted accordingly based on the original sample frequency including all the data.

4.4 High versus low liquidity assets

Estimators such as TSRV are designed to work for highly liquid assets. Indeed, the bias-correction relies on the idea that RV computed with all the observations, $[Y,Y]_T^{(\text{all})}$, consists primarily of noise: recall (15). This is of course true asymptotically in n. But if the full data sample frequency is low to begin with (for example, a stock sampled every minute instead of every second), $[Y,Y]_T^{(\text{all})}$ will not be entirely noise and bias-correcting on the basis of (15) may over-correct including in extreme cases possibly yielding a negative estimator in (16). So care may be taken to apply the estimator to settings which are appropriate: this is designed to work for very high frequency data, meaning settings where the raw data are sampled every few seconds in the case of typical financial data.

This number is of course to be assessed in relation to the magnitude of the noise in the raw data: the smaller the noise to begin with, the more frequent the observations. In particular, the estimator is not designed to be applied to data that have been subjected to preliminary de-noising steps such as MA(1) filters and the like. It is designed to work on the raw data, without requiring any step other than the correction of obvious data errors, such as prices entered as 0, etc. This is a strength of the approach in that it does not necessitate that one takes a stand on when and how to "pre-clean" the data and no other outside intervention (and the inevitable arbitrariness that come with them.)

4.5 Robustness to data cleaning procedures

We discuss in Aït-Sahalia et al. (2005) the importance and, often, the lack of robustness of results reported in the literature on specific data treatment procedures. There, we reproduced the analysis of Hansen and Lunde (2006) on a small subset of their data, for one stock (Alcoa, ticker symbol: AA) and one month (January 2004). We found that over half of the original dataset was discarded when creating a "clean" set of transactions. This set of manipulations in fact result in significant differences for the estimation, that may not be justified. For this particular stock and month, there were no transaction prices reported at 0 and this particular sample appears to be free of major data entry errors. Discarding all price moves of magnitude 0.5% or greater that are immediately followed by another move of the same size but the opposite sign, eliminated fewer than 10 observations each day. So a minimal amount of data cleaning would discard a very tiny percentage of the original transactions, nowhere near half of the sample.

The "clean" data are smoothed to the point where the estimator analyzed by Hansen and Lunde (2006) looks in fact very close to the basic uncorrected

RV and appears to underestimate the quadratic variation by about 20%. Finally, the data cleaning may significantly change the autocorrelation structure of returns. The "clean" dataset results in a first order autocorrelation (which is indicative of the inherent i.i.d. component of the noise of the data) that is about a *quarter* of the value obtained from the raw data. Therefore the main manifestation of the noise, namely the first order autocorrelation coefficient, has been substantially altered. At the same time, those cleaning procedures seem to introduce spuriously higher positive autocorrelation at orders 3 and above. So, at least for the data we analyzed, heavy-handed pre-processing of the data is far from being inconsequential.

4.6 Smoothing by averaging

Another issue is that the empirical analysis often conducted in the literature involves estimators computed on a daily basis, and then averaged over a longer time period, such as a year. While this time series averaging has the advantage of delivering plots that visually appear to be very smooth, it is not clear to us that this is how such estimators would be used in applications such as volatility hedging or option pricing. The whole point of using nonparametric measurements of *stochastic* volatility, estimated on a *day-by-day* basis, is that one believes that the quantity of interest can change meaningfully every day, at least for the purposes for which it is to be used (such as adjusting a position hedge). While some averaging is perhaps necessary, computing an average of the day-by-day numbers over an entire year seems to be at odds with the premise of the exercise.

One consequence of the large pre-processing of the data discussed above is that it reduces the sample size available for inference, which inevitably increases the variability of the estimators, i.e., decreases the precision with which we can estimate the quadratic variation. Examples of what happens when various estimators are implemented on a single day's data are reported in Aït-Sahalia et al. (2005b). We find that many estimators tend to be very sensitive to the frequency of estimation, such as RV implemented at, say, 4mn vs. 5mn vs. 6mn where the estimates turn out to be surprisingly different from those apparently small changes in what is essentially an arbitrary sampling frequency. Autocovariance-based estimators can also be quite sensitive to the number of lags included.

By contrast, the averaging *over sampling frequencies* that takes place in TSRV within a given trading day provides the necessary smoothing. This seems to us to be substantially better than having to average over different days, with all the implicit stationarity assumptions made in the process and the disconnect with the practical applications.

5 Conclusions

High frequency financial data have many unpleasant features from the perspective of econometric analysis. Any estimator that is robust to microstructure noise should attempt to address them, with minimal prior intervention by the researcher: once should not have to discard arbitrarily a large portion of the sample. Beyond eliminating the obvious data errors, such as prices or volumes entered as zeroes, one should not have to make up algorithms on the fly to determine which transactions are "better" than others, which quotes are valid or not, whether transactions time-stamped at the same second are truly contemporaneous or not, which market is more or less efficient at providing price discovery, etc. In the end, this is what market microstructure noise is all about!

Avoiding arbitrariness in implementation is an important consideration but at the same time, the choice between various sampling mechanisms may be dictated by the data at hand, by the importance placed on the rounding mechanism, or on the likelihood that the resulting price series looks like a martingale, on what we are going to do with that volatility estimate and what it is supposed to represent, etc. Ultimately, we are hoping for estimators that are robust to these considerations. We do not think we are quite there yet, but in a sense the fact that the results may differ when implemented on quotes vs. transactions, or on tick vs. calendar time sampling, are all different manifestations of the presence of the noise.

The estimators we have reviewed in this chapter provide the first step towards constructing volatility estimators that have good properties, including the basic requirement of consistency, in the presence of some form of market microstructure noise, and share additional robustness properties against deviations from the basic theoretical framework.

References

Aït-Sahalia, Y. (2002): Telling from discrete data whether the underlying continuous-time model is a diffusion. *Journal of Finance* **57**, 2075–2112.

Aït-Sahalia, Y. (2004): Disentangling diffusion from jumps. *Journal of Financial Economics* **74**, 487–528.

Aït-Sahalia, Y. and Jacod, J. (2004): Fisher's information for discretely sampled Lévy processes. *Technical Report, Princeton University and Université de Paris VI.*

Aït-Sahalia, Y. and Jacod, J. (2007): Volatility estimators for discretely sampled Lévy processes. *Annals of Statistics* **35**, 335–392.

Aït-Sahalia, Y. and Mancini, L. (2006): Out of sample forecasts of quadratic variation. *Journal of Econometrics* forthcoming.

Aït-Sahalia, Y., Mykland, P. A. and Zhang, L. (2005): Comment on "realized variance and market microstructure noise". *Journal of Business and Economic Statistics* **24**, 162–167.

Aït-Sahalia, Y., Mykland, P. A. and Zhang, L. (2005a): How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* **18**, 351–416.

Aït-Sahalia, Y., Mykland, P. A. and Zhang, L. (2005b): Ultra high frequency volatility estimation with dependent microstructure noise. *Technical Report, Princeton University*.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001): The distribution of exchange rate realized volatility. *Journal of the American Statistical Association* **96**, 42–55.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2006): Regular and modified kernel-based estimators of integrated variance: The case with independent noise. *Technical Report, Department of Mathematical Sciences, University of Aarhus*.

Barndorff-Nielsen, O. E. and Shephard. N. (2002): Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Ser. B* **64**, 253–280.

Barndorff-Nielsen, O. E. and Shephard, N. (2004): Power and bipower variation with stochastic volatility and jumps (with discussion). *Journal of Financial Econometrics* **2**, 1–48.

Choi, J. Y., Salandro, D. and Shastri, K. (1988): On the estimation of bid-ask spreads: Theory and evidence. *The Journal of Financial and Quantitative Analysis* **23**, 219–230.

French, K. and Roll, R. (1986): Stock return variances: The arrival of information and the reaction of traders. *Journal of Financial Economics* **17**, 5–26.

Gatheral, J.and Oomen, R. (2007): Zero-intelligence realized volatility estimation. *Technical Report*.

Gençay, R., Ballocchi, G., Dacorogna, M., Olsen, R. and Pictet, O. (2002): Real-time trading models and the statistical properties of foreign exchange rates. *International Economic Review* **43**, 463–491.

Gloter, A. (2000): *Estimation des paramètres d'une diffusion cachée*. Ph.D. thesis, Université de Marne-la-Vallée.

Gloter, A. and Jacod, J. (2000): Diffusions with measurement errors: I - local asymptotic normality and II - optimal estimators. *Technical Report, Université de Paris VI*.

Goncalves, S. and Meddahi, N. (2005): Bootstrapping realized volatility. *Technical Report, Université de Montréal*.

Griffin, J. and Oomen, R. (2006): Sampling returns for realized variance calculations: Tick time or transaction time? *Econometric Reviews*, forthcoming.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer, New York.

Hall, P. and Heyde, C. C. (1980): *Martingale Limit Theory and Its Application*. Academic Press, Boston.

Hamilton, J. D. (1995): *Time Series Analysis*. Princeton University Press, Princeton.

Hansen, P. R. and Lunde, A. (2006): Realized variance and market microstructure noise. *Journal of Business and Economic Statistics* forthcoming.

Harris, L. (1990): Statistical properties of the Roll serial covariance bid/ask spread estimator. *Journal of Finance* **45**, 579–590.

Jacod, J. (1994): Limit of random measures associated with the increments of a Brownian semimartingale. *Technical Report, Université de Paris VI*.

Jacod, J. and Protter, P. (1998): Asymptotic error distributions for the euler method for stochastic differential equations. *Annals of Probability* **26**, 267–307.

Jacod, J. and Shiryaev, A. N. (2003): *Limit Theorems for Stochastic Processes (2nd ed.)*. Springer-Verlag, New York.

Lee, S. Y. and Mykland, P. A. (2006): Jumps in financial markets: A new nonparametric test and jump dynamics. *Technical Report, Georgia Institute of Technology and The University of Chicago. Review of Financial Studies* to appear.

Li, Y. and Mykland, P. A. (2007): Are volatility estimators robust with respect to modeling assumptions? *Bernoulli* **13**, 601–622.

Mancini, C. (2001): Disentangling the jumps of the diffusion in a geometric jumping Brownian motion. *Giornale dell'Istituto Italiano degli Attuari* **LXIV**, 19–47.

McCullagh, P. and Nelder, J. (1989): *Generalized Linear Models (2nd ed.)*. Chapman and Hall, London.

Mykland, P. A. and Zhang, L. (2006): ANOVA for diffusions and Itô processes. *Annals of Statistics* **34**, 1931–1963.

Protter, P. (2004): *Stochastic Integration and Differential Equations: A New Approach (2nd ed.)*. Springer-Verlag, New York.

Roll, R. (1984): A simple model of the implicit bid-ask spread in an efficient market. *Journal of Finance* **39**, 1127–1139.

Wedderburn, R. (1974): Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439–447.

Xiu, D. (2008): Quasi-maximum likelihood estimation of misspecified stochastic volatility models. *Technical Report, Princeton University*.

Zhang, L. (2001): *From martingales to ANOVA: Implied and realized volatility*. Ph.D. thesis, The University of Chicago, Department of Statistics.

Zhang, L. (2006): Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli* **12**, 1019–1043.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005a): Edgeworth expansions for realized volatility and related estimators. *Technical Report, University of Illinois at Chicago*.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005b): A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* **100**, 1394–1411.

Zhou, B. (1996): High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics* **14**, 45–52.

Zumbach, G., Corsi, F. and Trapletti, A. (2002): Efficient estimation of volatility using high frequency data. *Technical Report, Olsen & Associates*.

Option Pricing

Jan Kallsen

Abstract This chapter reviews basic concepts of derivative pricing in financial mathematics. We distinguish market prices and individual values of a potential seller. We focus mainly on arbitrage theory. In addition, two hedging-based valuation approaches are discussed. The first relies on quadratic hedging whereas the second involves a first-order approximation to utility indifference prices.

1 Introduction

The valuation of derivative securities constitutes one of the main topics in modern mathematical finance. More generally, economic theory has considered the genesis of asset prices — whether they are derivatives or not — for a long time. We focus here on *relative* pricing, i.e. the valuation of options in relation to some underlying whose price is assumed to be given exogenously. In many cases it is obvious whether to consider an asset as underlying or derivative security. In others as e.g. in interest rate theory, there may exist more than one reasonable choice. But what do we mean by *option pricing*? We distinguish two possible interpretations. Depending on the application one may prefer one or the other.

In the following two sections we concentrate on *market prices* for options. They result from or are subject to supply and demand. General equilibrium theory has produced a large number of qualitative and quantitative results on market prices, based on assumptions concerning reasonable behaviour of investors, their preferences, beliefs and endowment (cf. Duffie (2001), Föllmer and Schied (2004)). However, their application in order to derive real prices

Jan Kallsen
Christian-Albrechts-Universität zu Kiel, Christian-Albrechts-Platz 4, D-24098 Kiel, Germany , e-mail: kallsen@matn.uni-kiel.de

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_26,

constitutes a daunting task. Even if the underlying economic assumptions hold, one typically lacks the information on investors' preferences etc. that are necessary to come up with concrete numbers.

Nevertheless, some useful statements on asset prices *relative to each other* can be made without too much prior knowledge. Harrison and Kreps (1979) shows that if a securities market is *viable* in the sense of general equilibrium theory, then discounted asset prices are martingales relative to some equivalent probability measure (called *equivalent martingale measure* or *EMM*). Subsequently, viability has been replaced by the simpler concept of *absence of arbitrage*, which does not involve detailed assumptions on investors' behaviour, cf. Harrison and Pliska (1981). If markets do not allow for riskless gains in a sense to be made precise, then there exists some EMM and vice versa. This so-called *fundamental theorem of asset pricing* has been studied intensively and in depth. For a thorough account of the theory cf. Delbaen and Schachermayer (2006). We discuss the use and limits of the arbitrage concept in Sections 2 and 3.

Instead of considering market prices one may also adapt the *individual investor's point of view*. Suppose you are in the position of a bank that is approached by some customer who wants to buy a certain contingent claim. At this point you must decide at which minimum price you are willing to make the deal. Moreover, you may wonder how to invest the premium you receive in exchange for the option. From a general economic theory point of view this individual price is easier to determine than the above market values. You need not know the preferences etc. of all market participants. It is enough to have an idea of your own endowment, preferences and beliefs. But do you? Even if you do not, arbitrage theory provides at least partial answers. In some lucky cases as e.g. in the Black-Scholes model the theory even tells you exactly what to do; in other situations it does not help much. Somewhat subjectively, we discuss two alternative approaches in the latter cases, namely quadratic hedging based valuation in Section 5 and utility indifference pricing in Section 6.

Many other suggestions have been made how to come up with option prices in so-called *incomplete markets* where arbitrage arguments fail to single out unique values. We do not attempt here to do justice to these very diverse approaches, which are based e.g. on economical, statistical, or mathematical assumptions and which are formulated in very different setups (cf. e.g. Kallsen (2002), Karatzas and Shreve (1998)).

2 Arbitrage Theory from a Market Perspective

Arbitrage theory relies on the key assumption that market prices do not allow for perfectly riskless profits. This in turn can be deduced from general economic principles. But it is often motivated by appealing to common

sense: should slight arbitrage opportunities occur, someone will exploit them immediately and make them disappear.

Let $S = (S^0, \ldots, S^d)$ signify the vector-valued *price process* of $d+1$ traded securities in the market and fix a time horizon $T > 0$. The underlying filtered probability space is denoted as $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, P)$. *Trading strategies* are predictable or in some sense *admissible* processes $\phi = (\phi^0, \ldots, \phi^d)$ representing the number of shares of the individual assets in the portfolio. The *value* of the portfolio at time t is

$$V_t(\phi) := \phi_t^\top S_t := \sum_{i=0}^{d} \phi_t^i S_t^i.$$

A trading strategy ϕ is called *self-financing* if

$$V(\phi) = V_0(\phi) + \phi \bullet S,$$

which means that no funds are added or withdrawn after inception of the portfolio. Here,

$$\phi \bullet S_t = \int_0^t \phi_s dS_s = \sum_{i=0}^{d} \int_0^t \phi_s^i dS_s^i$$

denotes the stochastic integral of ϕ with respect to S, which stands for the gains from trade between 0 and t. For discrete-time processes, it reduces to a sum

$$\phi \bullet S_t = \sum_{s=1}^{t} \sum_{i=0}^{d} \phi_s^i (S_s^i - S_{s-1}^i).$$

By $\widehat{S} := S/S^0$ we denote *discounted* securities prices, i.e. all prices are expressed as multiples of the *numeraire* security S^0. Accordingly, $\widehat{V}(\phi) := V(\phi)/S^0$ is the discounted value of a portfolio ϕ. The numeraire is often chosen particularly simple, e.g. as deterministic money-market account. Roughly speaking, an *arbitrage opportunity* signifies some self-financing strategy ϕ with initial value $V_0(\phi) = 0$ and such that the terminal wealth $V_T(\phi)$ is non-negative and positive with positive probability.

Ignoring nasty technical details, a key statement can be phrased as follows:

Theorem 1 (Fundamental theorem of asset pricing, FTAP) *There are no arbitrage opportunities if and only if there exists some probability measure $Q \sim P$ such that the discounted asset price process $\widehat{S} = S/S^0$ is a Q-martingale.*

Q is called *equivalent martingale measure (EMM)* and it may or may not be unique. Theorem 1 holds literally in discrete time (cf. Dalang et al. (1990), Harrison and Pliska (1981)). In continuous time the situation is less obvious. In order for some FTAP to hold one must modify the notions of arbitrage and EMM carefully. A key issue concerns the choice of admissible integrands

in the definition of self-financing trading strategies. If one allows for arbitrary S-integrable processes ϕ, then arbitrage opportunities exist in many reasonable models including the Black-Scholes setup. If one restricts the set of admissible strategies too much — e.g. to piecewise constant ones which are the only strategies that are practically feasible — then many interesting results from mathematical finance including the FTAP cease to hold. Delbaen and Schachermayer (1994, 1998) introduce the notion of *no free lunch with vanishing risk (NFLVR)* as a moderate extension of *no arbitrage (NA)*. Their notion of admissibility requires the discounted wealth $\widehat{V}(\phi)$ to be bounded from below. In this case the FTAP holds with equivalent *σ-martingale measures*, i.e. $\widehat{S}$ is a σ-martingale (a slight generalization of martingale or local martingale) relative to some probability measure $Q \sim P$.

This version of the FTAP may be the most prominent and satisfactory one but it is by no means the only one. If the notion of admissibility or arbitrage is relaxed, the dual set of probability measures in the FTAP narrows. One may e.g. consider riskless gains in the sense that some positive random variable can be approximated in L^2 by the payoff of piecewise constant strategies. From Théorème 2 in Stricker (1990) it follows that the absence of such *L^2-free lunches* is equivalent to the existence of some true EMM Q with square-integrable density dQ/dP.

In the following we do not make such fine distinctions. By FTAP we loosely refer to some statement as in Theorem 1 without worrying about precise definitions. In particular we carelessly disregard the difference between martingales and σ-martingales which often leads to conceptual and technical problems.

What are the implications of the FTAP for option prices? Suppose that S^2 denotes the price process of a European-style contingent claim on S^1, more specifically with terminal value $S^2_T = f(S^1_T)$ with some function f as e.g. $f(x) = (x - K)^+$. If the market is arbitrage-free, then $\widehat{S}^1, \widehat{S}^2$ are Q-martingales relative to some $Q \sim P$, which implies

$$S^2_t = S^0_t E_Q(f(S^1_T)/S^0_T|\mathcal{F}_t) \tag{1}$$

for the option price process and in particular

$$S^2_0 = S^0_0 E_Q(f(S^1_T)/S^0_T) \tag{2}$$

for the initial value. Recall that the law of the underlying processes (here: S^0, S^1) relative to P is assumed to be given exogenously, e.g. based on statistical inference. The problem is that we do not know the market's EMM or *pricing measure* Q needed to compute option prices S^2_t via (1). But if there exists only one probability measure $Q \sim P$ such that $\widehat{S}^1$ is a Q-martingale, then it must be the one we need. In this case (1) resp. (2) leads to a unique price. The prime example is of course the Black-Scholes model, where S^0 is a deterministic bank account and S^1 follows geometric Brownian motion.

If there exist more than one Q such that $\widehat{S}^1$ is a Q-martingale, (2) yields only an interval of possible initial values. Consider e.g. a discrete version of the Black-Scholes model, namely a discrete-time process $S^1_t, t = 0, 1, \ldots, T$ such that the log-returns $\log(S^1_t/S^1_{t-1})$ are i.i.d. Gaussian random variables. For simplicity we assume a constant numeraire $S^0_t = 1$, $t = 0, 1, \ldots, T$. It turns out that the interval of possible initial prices for a European call with strike K is given by $((S^1_0 - K)^+, S^1_0)$. This means that for an out-of-the-money option (i.e. $K < S^1_0$) any positive option price below the stock price is consistent with the absence of arbitrage. In other words, arbitrage theory hardly provides any useful information on option prices in this case. The same holds for many continuous-time stock price models of the form $S^1_t = S^1_0 \exp(L_t)$ with some jump-type Lévy process L rather than Brownian motion.

Sometimes it helps to consider a few derivatives as additional underlyings. Suppose that we model not only the stock but also a number of call options $S^2, \ldots, S^k$ on the stock exogenously by statistical methods. This reduces the set of possible pricing measures Q in the FTAP because now $\widehat{S}^i$, $i = 1, \ldots, k$ must be Q-martingales rather than only $\widehat{S}^1$. This modelling must be done very carefully in order to warrant absence of arbitrage, i.e. to have at least one EMM Q for $S^1, \ldots, S^k$ (cf. e.g. Jacod and Protter (2006)). An alternative approach is discussed in the following section.

Let us note in passing that different forms of the FTAP must be applied for futures contracts, securities paying dividends, American options, etc. (cf. e.g. Björk (2004), Kallsen and Kühn (2005)).

3 Martingale Modelling

As noted above, arbitrage theory yields only limited information in many, in particular discrete-time models. We discuss here a way out that enjoys much popularity in practice and is based on methods from statistics. The idea is as follows: if the theory does not tell us the true pricing measure Q, we ask the market, i.e. we make inference on Q by observing option prices in the real world.

More specifically, we proceed as follows. Suppose that S^1 is a stock and $S^2, \ldots, S^d$ are derivatives on the stock. The FTAP tells us that there exists some probability measure Q such that all discounted assets (the underlying S^1 as well as the derivatives $S^2, \ldots, S^d$) are Q-martingales. Unfortunately, we cannot make inference on Q by statistical methods because prices move according to the objective probability measure P.

But as in statistics we can start by postulating a particular parametric (or non-parametric) class of models. More precisely, we assume an explicit parametric expression for the dynamics of S^1 under Q. Note that the parameter vector θ_Q must be chosen such that $\widehat{S}^1$ is a Q-martingale. On the other hand, the dynamics of the options $S^2, \ldots, S^d$ need *not* be specified ex-

plicitly. Since we want Q to be an EMM for $S^1, S^2, \ldots, S^d$, the evolution of $S^2, \ldots, S^d$ and in particular the initial prices $S_0^2, \ldots, S_0^d$ follow as in (1) and (2). If $S^2, \ldots, S^d$ are traded on the market, then the parameter vector θ_Q should be chosen such that the theoretical prices computed as in (2) match the observed market prices. This corresponds to a moment estimator for θ_Q. The "estimation" of θ_Q by equating theoretical and observed option prices is commonly called *calibration*. If the dimension of θ_Q is small compared to the number of observed prices, some approximation as e.g. a least-squares fit will be necessary. In the non-parametric case, on the other hand, one may wish to rely on methods from non-parametric statistics in order to avoid very non-smooth or irregular solutions (cf. Belomestny and Reiß (2006), Cont and Tankov (2004)).

For calibration purposes one needs to compute option prices quickly. For standard options with European-style payoff $f(S_T^1)$ efficient algorithms can be based on *fast Fourier transform (FFT)* techniques if the characteristic function of $\log S_T^1$ under Q is known in closed form (cf. Carr and Madan (1999), Raible (2000)). Therefore models with explicit characteristic function come in handy for modelling financial data using option prices.

Strictly speaking, we should distinguish two situations. If one only wants to obtain information on the functional dependence of option prices on the underlying, it is enough to consider the martingale measure Q. There is basically no need to model the objective measure P as well. If, on the other hand, one wants to make statements also on "real" probabilities, quantiles, expectations, etc., one must model the underlying under both P and Q. Typically, one chooses the same parametric class for the dynamics of S^1 under both measures. The parameter set θ_P is obtained by statistical inference from past data whereas the corresponding parameters θ_Q are determined from option prices as explained above. Note that some parameters must coincide under P and Q in order to warrant the equivalence $Q \sim P$ stated in the FTAP. Consequently, statistical inference does in fact yield at least partial information on Q-parameters.

How do we know whether our postulated class of models for Q is appropriate? At first sight there seems to be no answer because statistical methods yield information on P but not on the theoretical creation Q. However, some evidence may be obtained again from option prices. If no choice of the parameter vector θ_Q yields theoretical option prices that are reasonably close to observed market prices, then the given class is obviously inappropriate.

How can the model be applied to options that are not yet traded? This is obvious from a mathematical point of view: if we assume Q to be the pricing measure of the FTAP for the whole market (including the new claim that is yet to be priced), we can determine its initial value as Q-expectation of the discounted payoff as in (2).

However, in practice we should be aware of two problems. Application of the calibrated model to new payoffs means extrapolation and is consequently to be taken with care. Cont (2006), Schoutens et al. (2005) compare models

that produce vastly different prices for exotic options even though they are calibrated successfully to the same large set of plain vanilla options in the first place.

Secondly, it is not clear what these extrapolated prices mean to the single investor. The fact that they are consistent with absence of arbitrage does not imply that the new options can be hedged well or involve only small risk. A more individual view on option pricing is discussed in the following sections.

4 Arbitrage Theory from an Individual Perspective

In the remaining three sections we take the investor's point of view. For ease of notation we assume a constant numeraire $S^0 = 1$ or, equivalently, we refer to discounted prices. Moreover, we denote the underlying stock simply by S rather than S^1 above. It is the only liquidly traded asset except the numeraire. Recall that we suppose the law of S to be known. The individual investor's problem can be phrased as follows: if a customer approaches you in order to buy a contingent claim with payoff $H = f(S_T)$ at time T, what option price shall you charge at time 0? And how do you invest this premium reasonably?

The answer to these questions depends generally on your preferences etc., in particular on your attitude towards risk. A minimum price cannot be determined unless you make up your mind in this respect. In some cases, however, arbitrage theory allows one to proceed without such information. No matter what your precise preferences are, you are probably willing to accept riskless gains. This has implications on the set of reasonable prices as the following result shows. It is stated more precisely and proved in El Karoui and Quenez (1995), Kramkov (1996).

Theorem 2 (Superreplication) *Let $H = f(S_T)$ denote the payoff of some contingent claim. Then*

$$\begin{aligned}\pi_{\text{high}} &:= \min\big\{\pi \in \mathbb{R} : \textit{There exists some self-financing strategy } \phi \\ &\qquad \textit{with initial value } V_0(\phi) = \pi \textit{ and terminal value } V_T(\phi) \geq H\big\} \\ &= \sup\big\{E_Q(H) : Q \textit{ EMM for } S\big\}\end{aligned}$$

and accordingly

$$\begin{aligned}\pi_{\text{low}} &:= \max\big\{\pi \in \mathbb{R} : \textit{There exists some self-financing strategy } \phi \\ &\qquad \textit{with initial value } V_0(\phi) = \pi \textit{ and terminal value } V_T(\phi) \leq H\big\} \\ &= \inf\big\{E_Q(H) : Q \textit{ EMM for } S\big\}.\end{aligned}$$

Why does this result have implications on the minimum option price you require from the investor? If you receive any amount $\pi \geq \pi_{\text{high}}$, then you can buy a self-financing strategy with terminal value $V_T(\phi) \geq H$, i.e. it meets all your obligations at time T. You do not face the risk of losing any money. Consequently, there is no reason for you not to accept at least π_{high} as option premium, maybe plus some fee for administrative purposes.

On the other hand, there is no reason to be satisfied with less than π_{low}. Rather than selling the option at a premium $\pi < \pi_{\text{low}}$, you should rather go short in a trading strategy with initial price π_{low} and terminal payoff $V_T(\phi) \leq H$ (i.e. you must pay at most H at time T). This deal on the stock market yields a higher profit than the suggested option trade with premium $\pi < \pi_{\text{low}}$, which means that you should not accept the latter. Consequently, reasonable premiums belong to the interval

$$\big[\inf\{E_Q(H) : Q \text{ EMM for } S\}, \sup\{E_Q(H) : Q \text{ EMM for } S\}\big], \tag{3}$$

which coincides essentially with the set of possible initial market prices based on the FTAP in Section 2.

If this interval reduces to a singleton, the option can be replicated perfectly, i.e. there is some self-financing ϕ satisfying $V_T(\phi) = H$. As a result you know exactly what to do in such Black-Scholes like cases. You charge the initial costs $V_0(\phi)$ (plus administration fee) as option premium, and you trade according to the hedge ϕ. This procedure removes your risk of losing money entirely, at least in theory.

On the other hand, arbitrage theory hardly helps in the other extreme cases discussed in Section 2, e.g. if the price interval for a European call equals $[0, S_0]$. No customer will accept to pay something close to the stock price $S_0 = \pi_{\text{high}}$ as premium for a call that is far out of the money. Therefore it seems misleading to call the difference $\pi_{\text{high}} - \pi_{\text{low}}$ bid-ask spread as is sometimes done in textbooks. It rather corresponds to an extreme upper bound for the real spread.

5 Quadratic Hedging

What can you do if the interval (3) is large and does not provide real guidance what premium to charge? A way out is to try and hedge the option as efficiently as possible and to require some compensation for the remaining unhedgable risk. For reasons of mathematical tractability we measure risk in terms of mean squared error. In order to hedge the option, you choose the self-financing strategy ϕ^* minimizing the risk

$$\epsilon^2(\phi) := E((V_T(\phi) - H)^2). \tag{4}$$

In so-called *complete* models the minimal risk is 0 because the option can be replicated by some ϕ^*. This in turn holds if and only if the interval (3) reduces to a singleton, namely $V_0(\phi^*)$. In general, however, the mean squared hedging error $\epsilon^2(\phi^*)$ does not vanish. A reasonable suggestion may be to charge $\pi := V_0(\phi^*) + \lambda\epsilon^2(\phi^*)$ (or $\pi := V_0(\phi^*) + \lambda\epsilon(\phi^*)$) as option premium, where $\lambda\epsilon^2(\phi^*)$ (resp. $\lambda\epsilon(\phi^*)$) compensates for the unhedgable risk and λ denotes a parameter chosen according to your personal risk aversion.

How can one determine $\phi^*, V_0(\phi^*), \epsilon^2(\phi^*)$? If S happens to be a martingale under the objective measure P, the solution can be expressed in terms of the *Galtchouk-Kunita-Watanabe (GKW) decomposition* of H relative to S (cf. Föllmer and Schied (1986)). To this end, let $V_t := E(H|\mathcal{F}_t)$ denote the martingale generated by the option payoff H. It can be decomposed as

$$V_t = V_0 + \phi \bullet S_t + L_t,$$

where L denotes some martingale that is *orthogonal* to S in the sense that LS is a martingale. This GKW decomposition yields the objects of interest, namely $\phi^* = \phi$, $V_0(\phi^*) = V_0$, $\epsilon^2(\phi^*) = E(L_T^2)$. More explicitly, we have

$$V_0 = E(H),$$

$$\phi_t = \frac{d\langle V, S\rangle_t}{d\langle S, S\rangle_t}$$

(in the sense that $\langle V, S\rangle = \phi \bullet \langle S, S\rangle$, where $\langle \cdot, \cdot \rangle$ denotes the *predictable covariation process* from stochastic calculus), and

$$E(L_T^2) = E(\langle V, V - \phi^* \bullet S\rangle_T).$$

If S fails to be a martingale, the derivation of $\phi^*, V_0(\phi^*), \epsilon^2(\phi^*)$ gets more involved but the GKW decomposition still plays a key role. We refer to Schweizer (2001) for an overview on the subject and to Černý and Kallsen (2007), Gourieroux et al. (1998), Rheinländer and Schweizer (1997), Schweizer (1994) for concrete formulas in various degrees of generality.

6 Utility Indifference Pricing

Valuation based on quadratic hedging is easy to understand and mathematically well tractable compared to other approaches. Its economic justification is less obvious. The minimization criterion (4) penalizes gains and losses alike. This is unreasonable from an economic point of view and occasionally leads to counterintuitive results. We want to discuss an economically better founded alternative.

Suppose that you are an investor with initial endowment v_0. Your preferences are modelled in terms of some increasing, strictly concave *utility function* $u : \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ quantifying the degree of happiness you derive from a certain amount of money. Your goal is to maximize the expected utility from terminal wealth $E(u(V_T(\phi)))$ over all self-financing strategies ϕ with initial value $V_0(\phi) = v_0$. Recall that the law of the underlying S is assumed to be known. If the customer offers you a premium π in exchange for the contingent claim H, you have two choices: either to accept the offer or to reject it. If you reject, your expected utility of terminal wealth will be

$$U_0 := \sup_{\phi} E(u(V_T(\phi))) \tag{5}$$

maximizing over all self-financing ϕ with initial value $V_0(\phi) = v_0$. If you accept the deal, it equals instead

$$U_H(\pi) := \sup_{\phi} E(u(V_T(\phi) + \pi - H)) \tag{6}$$

maximizing over the same set of strategies ϕ. Of course you will only accept the deal if it raises your utility, i.e. if $U_H(\pi) \geq U_0$. The lowest price π with this property is called *utility indifference price*, and it represents the minimum premium we are looking for (cf. Becherer (2003), M. Frittelli (2000), Mania and Schweizer (2005) for references). The utility indifference approach also tells you how to invest your money. You trade according to the optimal strategy ϕ in (6). Unfortunately, the utility indifference price is very hard to determine even for standard utility functions and in simple models with i.i.d. log-returns.

A reasonable way out for practical purposes is to consider a first-order approximation. This has been worked out recently in Mania and Schweizer (2005) for exponential utility and in Kramkov and Sîrbu (2006), Kramkov and Sîrbu (2006) for utility functions on $\mathbb{R}_+$, partly motivated by Henderson (2002), Kallsen (2002). Below we give a heuristic account of the results without worrying about mathematical assumptions and arguments needed to make the statements precise. For the latter we refer to Kramkov and Sîrbu (2006, 2006), Mania and Schweizer (2005).

In the following we focus on one of the following standard utility functions: $u(x) = \log(x)$, $u(x) = x^{1-p}/(1-p)$ for $p \in (0, \infty) \backslash \{1\}$, or $u(x) = 1 - \exp(-px)$ for $p > 0$. One should note that utility indifference prices are typically non-linear in the claim: if you sell two rather than one share of an option you require more than twice the premium because of your non-linear attitude towards risk.

We denote by $\pi(n)$ the utility indifference price per unit if the customer wants to buy n options. The optimal strategy ϕ in (6) corresponding to n options sold at $\pi(n)$ (i.e. for $U_{nH}(n\pi(n))$ is denoted by $\phi(n)$. We want to determine $\pi(n)$ and $\phi(n)$ approximately for small n. To this end, we assume

a smooth dependence

$$\pi(n) = \pi(0) + n\delta + o(n)$$

with constants $\pi(0), \delta$. The limiting price $\pi(0)$ for very small numbers of options is studied in Davis (1997), Karatzas and Kou (1996). The quantity δ is some kind of derivative of the option price relative to the number n that is to be sold. Similarly, we expand

$$\phi(n) = \phi^* + n\eta + o(n),$$

where ϕ^* denotes the optimal strategy in the pure investment problem (5) and η represents a hedging strategy per unit of H. The goal now is to derive formulas for $\pi(0)$, δ and η. It turns out that they can be expressed in terms of the GKW decomposition of the claim after some suitable measure and numeraire changes.

We start by considering the pure investment problem (5). For ϕ^* to be the optimizer, we need

$$E\big(u(v_0 + (\phi^* + \psi) \bullet S_T)\big) \leq E\big(u(v_0 + \phi^* \bullet S_T)\big) = E\big(u(V_T(\phi^*))\big)$$

for any competing strategy $\phi^* + \psi$. A first-order Taylor expansion for small ψ yields

$$E\big(u(v_0 + (\phi^* + \psi) \bullet S_T)\big) \approx E\big(u(V_T(\phi^*))\big) + E\big(u'(V_T(\phi^*))(\psi \bullet S_T)\big). \quad (7)$$

This is dominated by $E(u(V_T(\phi^*)))$ if and only if the last term in (7) is nonnegative for any ψ. Define a probability measure Q_0 in terms of its density

$$\frac{dQ_0}{dP} := \frac{u'(V_T(\phi^*))}{c_1}, \quad (8)$$

where $c_1 := E(u'(V_T(\phi^*)))$ denotes the normalizing constant. The optimality condition can be rewritten as $E_{Q_0}(\psi \bullet S_T) \leq 0$ for all ψ, which holds if and only if S is a Q_0-martingale.

Hence we have shown that an arbitrary candidate strategy ϕ^* maximizes (5) if and only if Q_0 defined as in (8) is an EMM for S. This criterion allows to determine ϕ^* explicitly in a number of concrete models (cf. e.g. Goll and Kallsen (2003, 2000)). Moreover, it can be shown that Q_0 minimizes some distance relative to P among all EMM's Q, namely the reverse relative entropy $E(\log(dP/dQ))$ for $u(x) = \log(x)$, the $L^{1-1/p}$-distance $E((dQ/dP)^{1-1/p})$ for $u(x) = x^{1-p}/(1-p)$ and the relative entropy $E_Q(\log(dQ/dP))$ for $u(x) = 1 - \exp(-px)$ (cf. Bellini and Frittelli (2002), Kallsen (2002)).

Now we turn to the optimization problem including n options which are sold for $\pi(n)$ each. More specifically, we seek to maximize

$$\begin{aligned} g(\eta) &:= E\big(u(V_T(\phi(n)) + n\pi(n) - nH)\big) \\ &= E\big(u(V_T(\phi^*) + n(\pi(0) + n\delta + (\eta + o(1)) \bullet S_T - H) + o(n^2))\big) \\ &= E\big(u(V_T(\phi^*))\big) + nE\big(u'(V_T(\phi^*))(\pi(0) + n\delta + (\eta + o(1)) \bullet S_T - H)\big) \\ &\quad + \frac{n^2}{2} E\big(u''(V_T(\phi^*))(\pi(0) + \eta \bullet S_T - H)^2\big) + o(n^2) \end{aligned}$$

Let us consider utility functions of logarithmic and power type first, in which case we have $u''(x) = -pu'(x)/x$ (with $p = 1$ in the case of logarithmic utility). Consequently,

$$\begin{aligned} g(\eta) = c_0 &+ nc_1 E_{Q_0}\Big(\pi(0) + n\delta + (\eta + o(1)) \bullet S_T - H\Big) \\ &- n^2 c_1 \frac{p}{2} E_{Q_0}\left(\frac{V_T(\phi^*)}{v_0^2}\Big(\frac{\pi(0) + \eta \bullet S_T - H}{v_0^{-1} V_T(\phi^*)}\Big)^2\right) + o(n^2) \end{aligned}$$

with $c_0 := E(u(V_T(\phi^*)))$. Since Q_0 is an EMM, we have

$$E_{Q_0}((\eta + o(1)) \bullet S_T) = 0.$$

Now define $Q_{\phi^*} \sim Q_0$ via

$$\frac{dQ_{\phi^*}}{dQ_0} := \frac{V_T(\phi^*)}{v_0},$$

where the normalizing constant $E_{Q_0}(V_T(\phi^*)) = v_0$ coincides with the initial endowment. Since Q_0 is an EMM, we have that Q_{ϕ^*} is an EMM relative to the numeraire $V(\phi^*)$ or equivalently $V(\phi^*)/v_0$. In other words, $\widetilde{S} := Sv_0/V(\phi^*)$ is a Q_{ϕ^*}-martingale. Define discounted values relative to this numeraire $V(\phi^*)/v_0$ by $\widetilde{\pi}(0) := \pi(0)v_0/V(\phi^*)$ and $\widetilde{H} := Hv_0/V_T(\phi^*)$. We prefer the numeraire $V(\phi^*)/v_0$ to $V(\phi^*)$ because it does not depend on v_0 for power and logarithmic utility. Since η is considered to be self-financing, we have

$$\frac{\pi(0) + \eta \bullet S_T - H}{v_0^{-1} V_T(\phi^*)} = \widetilde{\pi}(0) + \eta \bullet \widetilde{S}_T - \widetilde{H}$$

(cf. Goll and Kallsen (2000), Prop. 2.1). This yields

$$g(\eta) = c_0 + nc_1\big(\pi(0) - E_{Q_0}(H)\big) + n^2 c_1 \left(\delta - \frac{p}{2v_0}\epsilon^2(\eta)\right) + o(n^2)$$

with

$$\epsilon^2(\eta) := E_{Q_{\phi^*}}\Big((\widetilde{\pi}(0) + \eta \bullet \widetilde{S}_T - \widetilde{H})^2\Big).$$

This is to be maximized as a function of η. By disregarding the $o(n^2)$-term, we find that η is the integrand in the GKW decomposition of the Q_{ϕ^*}-martingale $\widetilde{V}_t := E_{Q_{\phi^*}}(\widetilde{H}|\mathcal{F}_t)$ relative to $\widetilde{S}$, i.e.

$$\eta_t = \frac{d\langle \widetilde{V}, \widetilde{S}\rangle_t^{Q_{\phi^*}}}{d\langle \widetilde{S}, \widetilde{S}\rangle_t^{Q_{\phi^*}}}.$$

What about $\pi(0)$ and δ? The indifference criterion is $E(u(V_T(\phi^*))) = g(\eta)$, i.e.

$$c_0 = c_0 + nc_1\big(\pi(0) - E_{Q_0}(H)\big) + n^2 c_1 \left(\delta - \frac{p}{2v_0}\epsilon^2(\eta)\right) + o(n^2).$$

This implies

$$\pi(0) = E_{Q_0}(H),$$

which is equivalent to $\widetilde{\pi}(0) = E_{Q_{\phi^*}}(\widetilde{H})$. Moreover,

$$\delta = \frac{p}{2v_0}\epsilon^2(\eta) = \frac{p}{2v_0} E_{Q_{\phi^*}}\left(\langle \widetilde{V}, \widetilde{V} - \eta \bullet \widetilde{S}\rangle_T^{Q_{\phi^*}}\right),$$

where $\epsilon^2(\eta)$ can be interpreted as minimal expected squared hedging error if $\widetilde{H}$ is hedged with $\widetilde{S}$ relative to Q_{ϕ^*} (cf. Section 5).

In the case of exponential utility $u(x) = 1 - \exp(-px)$ we have $u''(x) = -pu'(x)$. Up to a missing numeraire change, this leads to the same results as above. Instead of the GKW decomposition of $\widetilde{V}$ relative to $\widetilde{S}$ under Q_{ϕ^*} we must consider the GKW decomposition of $V_t = E_{Q_0}(H|\mathcal{F}_t)$ relative to S under the EMM Q_0. Similarly as before we have

$$\pi(0) = E_{Q_0}(H),$$

$$\eta_t = \frac{d\langle V, S\rangle_t^{Q_0}}{d\langle S, S\rangle_t^{Q_0}},$$

$$\delta = \frac{p}{2}\epsilon^2(\eta),$$

where

$$\epsilon^2(\eta) = E_{Q_0}\left(\langle V, V - \eta \bullet S\rangle_T^{Q_0}\right)$$

now corresponds to the minimal expected squared hedging error if H is hedged with S relative to Q_0.

References

Becherer, D. (2003): Rational hedging and valuation of integrated risks under constant absolute risk aversion. *Insurance: Mathematics and Economics* **33**, 1–28.

Bellini, F. and Frittelli, M. (2002): On the existence of minimax martingale measures. *Mathematical Finance* **12**, 1–21.

Belomestny, D. and Reiß, M. (2006): Spectral calibration of exponential Lévy models. *Finance & Stochastics* **10**, 449–474.

Björk, T. (2004): *Arbitrage Theory in Continuous Time*. 2nd edition. Oxford University Press, Oxford.

Carr, P. and Madan, D. (1999): Option valuation using the fast Fourier transform. *The Journal of Computational Finance* **2**, 61–73.

Černý, A. and Kallsen, J. (2007): On the structure of general mean-variance hedging strategies. *The Annals of Probability* to appear.

Cont, R. (2006): Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance* **16**, 519–547.

Cont, R. and Tankov, P. (2004): *Financial Modelling with Jump Processes*. Chapman & Hall/CRC, Boca Raton.

Dalang, R., Morton, A. and Willinger, W. (1990): Equivalent martingale measures and no-arbitrage in stochastic security market models. *Stochastics and Stochastics Reports* **29**, 185–202.

Davis, M. (1997): Option pricing in incomplete markets. *In: Dempster, M. and Pliska, S. (Eds.): Mathematics of Derivative Securities*, 216–226. Cambridge University Press, Cambridge.

Delbaen, F. and Schachermayer, W. (1994): A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* **300**, 463–520.

Delbaen, F. and Schachermayer, W. (1998): The fundamental theorem of asset pricing for unbounded stochastic processes. *Mathematische Annalen* **312**, 215–250.

Delbaen, F. and Schachermayer, W. (2006): *The Mathematics of Arbitrage*. Springer, Berlin.

Duffie, D. (2001): *Dynamic Asset Pricing Theory*. 3rd edition. Princeton University Press, Princeton.

El Karoui, N. and Quenez, M. (1995): Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM Journal on Control and Optimization* **33**, 29–66.

Föllmer, H. and Schied, A. (2004): *Stochastic Finance: An Introduction in Discrete Time*. 2nd edition. Walter de Gruyter, Berlin.

Föllmer, H. and Sondermann, D. (1986): Hedging of nonredundant contingent claims. *In: Hildenbrand, W. and Mas-Colell, A. (Eds.): Contributions to Mathematical Economics*, 205–223. North-Holland, Amsterdam.

Frittelli, M. (2000): Introduction to a theory of value coherent with the no-arbitrage principle. *Finance & Stochastics* **4**, 275–297.

Goll, T. and Kallsen, J. (2000): Optimal portfolios for logarithmic utility. *Stochastic Processes and their Applications* **89**, 31–48.

Goll, T. and Kallsen, J. (2003): A complete explicit solution to the log-optimal portfolio problem. *The Annals of Applied Probability*, **13**, 774–799.

Gourieroux, C., Laurent, J. and Pham, H. (1998): Mean-variance hedging and numéraire. *Mathematical Finance* **8**, 179–200.

Harrison, M. and Kreps, D. (1979): Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* **20**, 381–408.

Harrison, M. and Pliska, S. (1981): Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications* **11**, 215–260.

Henderson, V. (2002): Valuation of claims on non-traded assets using utility maximization. *Mathematical Finance* **12**, 351–371.

Jacod, J. and Protter, P. (2006): *Risk neutral compatibility with option prices. Preprint, 2006.*

Kallsen, J. (2000): Optimal portfolios for exponential Lévy processes. *Mathematical Methods of Operations Research* **51**, 357–374.

Kallsen, J. (2002): Derivative pricing based on local utility maximization. *Finance & Stochastics* **6**, 115–140.

Kallsen, J. (2002): Utility-based derivative pricing in incomplete markets. *In: Geman, H., Madan, D., Pliska, S. and Vorst, T. (Eds.): Mathematical Finance – Bachelier Congress 2000*, 313–338. Berlin, Springer.

Kallsen, J. and Kühn, C. (2005): Convertible bonds: Financial derivatives of game type. *In: Kyprianou, A., Schoutens, W. and Wilmott, P. (Eds.): Exotic Option Pricing and Advanced Lévy Models*, 277–291. Wiley, New York.

Karatzas, I. and Kou, S. (1996): On the pricing of contingent claims under constraints. *The Annals of Applied Probability* **6**, 321–369.

Karatzas, I. and Shreve, S. (1998): *Methods of Mathematical Finance*. Springer, Berlin.

Kramkov, D. (1996): Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets. *Probability Theory and Related Fields* **105**, 459–479.

Kramkov, D. and Sîrbu, M. (2006): Asymptotic analysis of utility based hedging strategies for a small number of contingent claims. *Preprint, 2006*.

Kramkov, D. and Sîrbu, M. (2006): The sensitivity analysis of utility based prices and the risk-tolerance wealth processes. *The Annals of Applied Probability* **16**, 2140–2194.

Mania, M. and Schweizer, M. (2005): Dynamic exponential utility indifference valuation. *The Annals of Applied Probability* **15**, 2113–2143.

Raible, S. (2000): *Lévy Processes in Finance: Theory, Numerics, and Empirical Facts. PhD thesis, University of Freiburg*.

Rheinländer, T. and Schweizer, M. (1997): On L^2-projections on a space of stochastic integrals. *The Annals of Probability* **25**, 1810–1831.

Schoutens, W., Simons, E. and Tistaert, J. (2005): Model risk for exotic and moment derivatives. *In: Kyprianou, A., Schoutens, W. and Wilmott, P. (Eds.): Exotic Option Pricing and Advanced Lévy Models*, 67–97. Wiley, New York.

Schweizer, M. (1994): Approximating random variables by stochastic integrals. *The Annals of Probability* **22**, 1536–1575.

Schweizer, M. (2001): A guided tour through quadratic hedging approaches. *In: Jouini, E., Cvitanic, J. and Musiela, M. (Eds.): Option Pricing, Interest Rates and Risk Management*, 538–574. Cambridge University Press, Cambridge.

Stricker, C. (1990): Arbitrage et lois de martingale. *Annales de l'Institut Henri Poincaré* **26**, 451–460.

An Overview of Interest Rate Theory

Tomas Björk

Abstract In this paper we give a short overview of some basic topics in interest rate theory, from the point of view of arbitrage free pricing. We cover short rate models, affine term structure models, inversion of the yield curve, the Musiela parameterization, and the potential approach to positive interest rates. The text is essentially self contained.

1 General Background

We consider a financial market model on a finite time interval $[0, \widehat{T}]$ living on a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, P)$ where $\mathbf{F} = \{\mathcal{F}_t\}_{t \geq 0}$ and P is interpreted as the "objective" or "physical" probability measure. The basis is assumed to carry a standard m-dimensional Wiener process W, and we also assume that the filtration $\mathbf{F}$ is the internal one generated by W. The choice of a Wiener filtration is made for convenience, and the theory below can be extended to a general semimartingale framework.

We assume that there exist $N+1$ non-dividend-paying assets on the market, and the prices at time t of these assets are denoted by $S_0(t), S_1(t), \cdots, S_N(t)$. We assume that the price processes are Itô processes and that $S_0(t) > 0$ with probability one. We view the price vector process $S = (S_0, S_1, \cdots S_N)^*$ as a column vector process, where * denotes transpose.

A **portfolio** is any adapted (row vector) process $h = (h^0, h^1, \cdots, h^N)$, where we interpret h_t^i as the number of units that we hold of asset i in the portfolio at time t. The corresponding market **value** process V^h is defined by $V^h(t) = h(t)S(t) = \sum_{i=0}^{N} h^i(t)S_i(t)$, and the portfolio is said to be **self financing** if the condition $dV(t) = h(t)dS(t)$ is satisfied.

Thomas Björk
Department of Finance, Stockholm School of Economics, P.O. Box 6501, S-113 83 Stockholm, SWEDEN, e-mail: tomas.bjork@hhs.se

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_27,

An **arbitrage** possibility is a self financing portfolio h with the properties that $V^h(0) = 0$, $P(V^h(T) \geq 0) = 1$ and $P(V^h(T) > 0) > 0$. An arbitrage would constitute a "money-making machine" and a minimal requirement of market efficiency is that the market is free of arbitrage possibilities. The main result in this direction is, subject to some technical conditions, as follows.

Theorem 1 *The market is free of arbitrage if and only if there exists a probability measure Q with the properties that*

1. $Q \sim P$
2. *All* **normalized** *asset processes*

$$\frac{S_0(t)}{S_0(t)}, \frac{S_1(t)}{S_0(t)}, \ldots, \frac{S_N(t)}{S_0(t)}$$

are Q-martingales.

Such a measure Q (which is typically not unique, see below) is called a **martingale measure**. The **numeraire asset** S_0 could in principle be any asset with positive prices, but very often it is chosen as the **money account** B defined by $dB(t) = r(t)B(t)dt$ where r is the short interest rate, i.e.,

$$B(t) = e^{\int_0^t r(s)ds}.$$

A **contingent T-claim** is any random variable $X \in \mathcal{F}_T$, where the interpretation is that the holder of the claim will receive the stochastic amount X (in a given currency) at time T. Given a T-claim X, a self financing portfolio h is said to **replicate** (or "hedge against") X if $V^h(T) = X$, P-a.s. The market model is **complete** if every claim can be replicated. The main result for completeness in an arbitrage free market is the following.

Theorem 2 *The market is complete if and only if the martingale measure is unique.*

We now turn to the pricing problem for contingent claims. In order to do this, we consider the "primary" market $S_0, S_1, \ldots, S_N$ as given *a priori*, and we fix a T-claim X. Our task is that of determining a "reasonable" price process $\Pi(t; X)$ for X, and we assume that the primary market is arbitrage free. There are two main approaches:

- The derivative should be priced in a way that is **consistent** with the prices of the underlying assets. More precisely we should demand that the extended market $\Pi(t; X), S_0(t), S_1(t), \ldots, S_N(t)$ is free of arbitrage possibilities.
- If the claim is **attainable**, with hedging portfolio h, then the only reasonable price is given by $\Pi(t; X) = V(t; h)$.

In the first approach above, we thus demand that there should exist a martingale measure Q for the extended market $\Pi(t; X), S_0(t), S_1(t), \ldots, S_N(t)$.

Letting Q denote such a measure, assuming enough integrability, and applying the definition of a martingale measure we obtain

$$\frac{\Pi(t;X)}{S_0(t)} = E^Q\left[\frac{\Pi(T;X)}{S_0(T)}\middle|\mathcal{F}_t\right] = E^Q\left[\frac{X}{S_0(T)}\middle|\mathcal{F}_t\right].$$

We thus have the following result.

Theorem 3 (General Pricing Formula) *The arbitrage free price process for the T-claim X is given by*

$$\Pi(t;X) = S_0(t)E^Q\left[\frac{X}{S_0(T)}\middle|\mathcal{F}_t\right], \tag{1}$$

where Q is the (not necessarily unique) martingale measure for the a priori given market $S_0, S_1, \cdots, S_N$, with S_0 as the numeraire .

Note that different choices of Q will generically give rise to different price processes.

In particular we note that if we assume that if S_0 is the money account

$$S_0(t) = S_0(0)\cdot e^{\int_0^t r(s)ds},$$

where r is the short rate, then (1) reduces to the familiar "risk neutral valuation formula".

Theorem 4 (Risk Neutral Valuation Formula)
Assuming the existence of a short rate, the pricing formula takes the form

$$\Pi(t;X) = E^Q\left[e^{-\int_t^T r(s)ds}X\middle|\mathcal{F}_t\right].$$

where Q is a (not necessarily unique) martingale measure with the money account as the numeraire.

For the second approach to pricing let us assume that X can be replicated by h. Since the holding of the derivative contract and the holding of the replicating portfolio are equivalent from a financial point of view, we see that the price of the derivative must be given by the formula

$$\Pi(t;X) = V^h(t). \tag{2}$$

One problem here is what will happen in a case when X can be replicated by two different portfolios, and one would also like to know how this formula is connected to (1).

Defining $\Pi(t;X)$ by (2) we note that the process $\Pi(t;X)/S_0(t)$ is a normalized asset price and thus a Q-martingale. Consequently we again obtain the formula (1) and for an attainable claim we have in particular the formula

$$V^h(t) = S_0(t)E^Q\left[\frac{X}{S_0(T)}\middle|\mathcal{F}_t\right],$$

which will hold for any replicating portfolio and for any martingale measure Q. Thus we see that the two pricing approaches above do in fact coincide on the set of attainable claims.

We finish with a remark on the characterization of a risk neutral martingale measure.

Lemma 1 *A risk neutral martingale measure, i.e., an EMM with the bank account as numeraire, is characterized by the properties that $Q \sim P$, and that every asset price process has the short rate as its local rate of return under Q. More precisely, under Q the dynamics of any asset price process π (derivative or underlying) must be of the form*

$$d\pi_t = \pi_t r_t dt + \pi_t \sigma_t^\pi dW_t^Q, \tag{3}$$

where r is the short rate and W^Q is Q-Wiener.

2 Interest Rates and the Bond Market

Our main object of study is the zero coupon bond market, and we need some formal definitions.

Definition 1 A **zero coupon bond** with **maturity date** T, also called a T-bond, is a contract which guarantees the holder \$1 to be paid on the date T. The price at time t of a bond with maturity date T is denoted by $p(t,T)$.

Given the bond market above, one can define a (surprisingly large) number of **riskless interest rates**. The term LIBOR below, is an acronym for "London Interbank Offer Rate".

Definition 2 1. The **simple forward rate for $[\mathbf{S},\mathbf{T}]$ contracted at t**, often referred to as the **LIBOR** forward rate, is defined as

$$L(t;S,T) = -\frac{p(t,T)-p(t,S)}{(T-S)p(t,T)}.$$

2. The **simple spot rate for $[\mathbf{S},\mathbf{T}]$**, henceforth referred to as the **LIBOR** spot rate, is defined as

$$L(S,T) = -\frac{p(S,T)-1}{(T-S)p(S,T)}.$$

3. The **continuously compounded forward rate for** $[\mathbf{S}, \mathbf{T}]$ **contracted at t** is defined as

$$R(t; S, T) = -\frac{\log p(t, T) - \log p(t, S)}{T - S}.$$

4. The **continuously compounded spot rate, for** $[\mathbf{S}, \mathbf{T}]$ is defined as

$$R(S, T) = -\frac{\log p(S, T)}{T - S}.$$

5. The **instantaneous forward rate with maturity T, contracted at t,** is defined by

$$f(t, T) = -\frac{\partial \log p(t, T)}{\partial T}.$$

6. The **instantaneous short rate at time t** is defined by

$$r(t) = f(t, t).$$

We now go on to define the money account process B.

Definition 3 The **money account** process is defined by

$$B(t) = e^{\int_0^t r(s)ds},$$

i.e.,

$$dB(t) = r(t)B(t)dt, \quad B(0) = 1.$$

The interpretation of the money account is that you may think of it as describing a bank with the stochastic short rate r.

As an immediate consequence of the definitions we have the following useful formulas.

Lemma 2 *For $t \leq s \leq T$ we have*

$$p(t, T) = p(t, s) \cdot e^{-\int_s^T f(t,u)du},$$

and in particular

$$p(t, T) = e^{-\int_t^T f(t,u)du}.$$

We finish this section by presenting the relations that hold between the dynamics of forward rates and those of the corresponding bond prices. These relations will be used repeatedly below. We will consider dynamics of the following form.

Bond price dynamics

$$dp(t, T) = p(t, T)m(t, T)dt + p(t, T)v(t, T)dW(t). \tag{4}$$

Forward rate dynamics

$$df(t,T) = \alpha(t,T)dt + \sigma(t,T)dW(t). \tag{5}$$

The Wiener process W is allowed to be vector valued, in which case the volatilities $v(t,T)$ and $\sigma(t,T)$ are row vectors. The processes $m(t,T)$, $v(t,T)$, $\alpha(t,T)$ and $\sigma(t,T)$ are allowed to be arbitrary adapted processes parameterized by time of maturity T.

Our main technical tool is as follows. The proof is omitted.

Proposition 1 *If $f(t,T)$ satisfies (5) then $p(t,T)$ satisfies*

$$dp(t,T) = p(t,T)\left\{r(t) + A(t,T) + \frac{1}{2}\|S(t,T)\|^2\right\}dt + p(t,T)S(t,T)dW(t),$$

where $\|\cdot\|$ denotes the Euclidean norm, and

$$\begin{cases} A(t,T) = -\int_t^T \alpha(t,s)ds, \\ \\ S(t,T) = -\int_t^T \sigma(t,s)ds. \end{cases}$$

3 Factor Models

Since the price of a contingent claim Z is given by the general formula

$$\Pi\,(t;Z) = E^Q\left[e^{-\int_t^T r_s ds}Z \;\middle|\; \mathcal{F}_t\right],$$

it is natural to study Markovian factor models of the form

$$\begin{aligned} dX_t &= \mu(t,X_t)dt + \sigma(t,X_t)dW_t, \\ r_t &= h(t,X_t), \end{aligned}$$

where μ , σ, and h are given deterministic functions and W is Wiener. In this framework we typically restrict ourselves to contingent T-claims Z of the form $Z = \Phi(X_T)$, where Φ denotes the **contract function**, i.e., Φ specifies the amount of money to be paid to the holder of the contract at time T. This modeling can be done either under the objective measure P, or under a martingale measure Q.

We recall that the defining properties of a risk neutral martingale measure Q are that $Q \sim P$ and that Π_t/B_t should be a Q-martingale for every asset price process Π. Since, in the present setup, the only asset price specified a priori is the bank account B_t, and since $B_t/B_t = 1$ is trivially a Q-martingale, we see that in this case **every** measure $Q \sim P$ is a martingale measure, and

that a particular choice of Q will generate arbitrage free asset prices by the prescription

$$\Pi(t;Z) = E^Q\left[e^{-\int_t^T h(s,X_s)ds}\Phi(X_T)\middle|\mathcal{F}_t\right], \tag{6}$$

for any claim Z of the form $Z = \Phi(X_T)$.

4 Modeling under the Objective Measure P

As above we consider a factor model of the form

$$\begin{aligned} dX_t &= \mu^p(t,X_t)dt + \sigma(t,X_t)dW_t^p, \\ r_t &= h(t,X_t), \end{aligned}$$

where W^p is P-Wiener. The price of claim Z of the form $Z = \Phi(X_T)$ is again given by the formula (6) above. In the present setting, with the filtration generated by W^p, it follows that the likelihood process L, defined by

$$L_t = \frac{dQ}{dP}, \quad \text{on } \mathcal{F}_t$$

is obtained by a Girsanov transformation of the form

$$dL_t = L_t\varphi_t^\star dW_t^p, \quad L_0 = 1,$$

where * denotes transpose. To keep the Markovian structure we now assume that the Girsanov kernel process φ is of the form $\varphi(t,X_t)$ and from the Girsanov Theorem we can write $dW_t^p = \varphi_t dt + dW_t$ where W is Q-Wiener. We thus have the Q-dynamics of X as

$$dX_t = \{\mu^p(t,X_t) + \sigma(t,X_t)\varphi(t,X_t)\}\,dt + \sigma(t,X_t)dW_t.$$

For notational simplicity we denote the Q-drift of X by μ, i.e.,

$$\mu(t,x) = \mu^p(t,x) + \sigma(t,x)\varphi(t,x).$$

Since the price process $\Pi(t;Z)$ for a claim of the form $Z = \Phi(X_T)$ is given by (6) we now have the following result, which follows directly from the Kolmogorov backward equation.

Theorem 5

- *For a claim of the form $Z = \Phi(X_T)$, the price process $\Pi(t;Z)$ is in fact of the form $\Pi(t;Z) = F(t,X_t)$ where F satisfies the* **term structure equation**

$$\frac{\partial F}{\partial t}(t,x) + \mathcal{A}F(t,x) - h(t,x)F(t,x) = 0, \tag{7}$$

$$F(T,x) = \Phi(x), \tag{8}$$

where the operator $\mathcal{A}$ is given by

$$\mathcal{A}F(t,x) = \sum_{i=1}^{n} \mu_i(t,x)\frac{\partial F}{\partial x_i}(t,x) + \frac{1}{2}\sum_{i,j=1}^{n} C_{ij}(t,x)\frac{\partial^2 F}{\partial x_i \partial x_j}(t,x),$$

and where $C(t,x) = \sigma(t,x)\sigma^\star(t,x)$.

- *In particular, bond prices are given by $p(t,T) = F^T(t,X_t)$ (the index T is viewed as a parameter), where the pricing function F^T satisfies*

$$\frac{\partial F^T}{\partial t}(t,x) + \mathcal{A}F^T(t,x) - h(t,x)F^T(t,x) = 0, \tag{9}$$

$$F^T(T,x) = 1. \tag{10}$$

4.1 The market price of risk

There is an immediate economic interpretation of the Girsanov kernel φ above. To see this let π_t be the price process of any asset (derivative or underlying) in the model. We write the P-dynamics of π as

$$d\pi_t = \pi_t \alpha_t dt + \pi_t \delta_t dW_t^P,$$

where α is the local mean rate of return of π (under P) and δ is the (vector) volatility process. From the Girsanov Theorem we obtain, as above,

$$d\pi_t = \pi(t)\left\{\alpha_t + \delta_t\varphi_t\right\}dt + \pi_t\delta_t dW_t,$$

where W is Q-Wiener. From Lemma 3 we have , on the other hand,

$$d\pi_t = \pi_t r_t dt + \pi_t \delta_t dW_t,$$

so we obtain the relation

$$\alpha_t + \delta_t\varphi_t = r_t,$$

or, equivalently,

$$\alpha_t - r_t = \delta_t\varphi_t = -\sum_i \delta_{it}\varphi_{it}.$$

In other words, the **risk premium** for π, given by $\alpha_t - r_t$, i.e., the excess rate of return above the risk free rate r, is given (apart from a minus sign) as the sum of the volatility terms δ_i multiplied by the "factor loadings" δ_i. This has motivated economists to refer to the process $\lambda_t = -\varphi_t$ as the "market price of risk" process, where λ_i is the market price of risk for Wiener factor number i. In particular we see that if W (and thus δ) is scalar then λ in fact equals the **Sharpe ratio**, i.e.,

$$\lambda_t = \frac{\alpha_t - r_t}{\delta_t}.$$

The economic interpretation is that λ is a measure of the aggregate risk aversion in the market, in the sense that if λ is positive then the market is risk averse, if λ is negative then the market is risk loving and if $\lambda = 0$ is positive then the market is risk neutral. We summarize the moral in the following slogan.

Result 1 *The martingale measure is chosen by* **the market**.

5 Martingale Modeling

In order to construct a factor model of the type above, and to be able to compute derivative prices, it seems that we have to model the following objects.

- The P-drift μ^p.
- The volatility σ (which is the same under P and under Q.
- The market price of risk $\lambda = -\varphi$, which connects Q to P by a Girsanov transformation.

However, from the pricing formula (6) we have the following simple observation.

Proposition 2 *The term structure of bond prices, as well as the prices of all other derivatives, are completely determined by* **specifying the dynamics of X under the martingale measure Q**.

This observation has led to the following standard modeling procedure: instead of specifying μ^p, σ and λ under the objective probability measure P we will henceforth specify the dynamics of the factor process X directly under the martingale measure Q. This procedure is known as **martingale modeling**, and the typical assumption will thus be that X under Q has dynamics given by

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{11}$$

where W is Q-Wiener.The short rate is as before defined by

$$r_t = h(t, X_t). \tag{12}$$

The pricing formulas from Theorem 5 still hold.

5.1 Affine term structures

In order to compute prices of derivatives, we have to be able so solve the term structure equation (7)-(8). It turns out that the only cases when the term structure equation can be solved analytically is more or less when we have an **affine term structure**.

Definition 4 The factor model (11) -(12) above is said to possess an affine term structure (AFT for short) if bond prices are of the form

$$p(t,T) = e^{A(t,T)-B(t,T)X_t}, \tag{13}$$

where A (scalar) and B (row vector) are deterministic functions of t and T.

The importance of the AFT models stem from the fact that these are roughly speaking the only models for which we can obtain analytical formulas for bond prices and bond option prices. The question now arises as to when we have an AFT, and the basic result is as follows.

Theorem 6 *Sufficient conditions for the existence of an affine term structure are the following.*

1. *The drift (under Q) is an affine function of the factors, i.e., μ is of the form*
$$\mu(t,x) = \alpha(t) + \Delta(t)x,$$
where the $n \times 1$ column vector α and the $n \times n$ matrix Δ are deterministic functions of time.
2. *The "square of the diffusion" is an affine function of the factors, i.e., $\sigma\sigma^\star$ is of the form*
$$\sigma(t,x)\sigma^\star(t,x) = C(t) + \sum_{i=1}^{n} D_i(t)x_i,$$
where C and D_i are deterministic $n \times n$ matrix functions of t.
3. *The short rate is an affine function of the factors, i.e.,*
$$h(t,x) = c(t) + d(t)x,$$
where the scalar c and the $1 \times n$ row vector d are deterministic functions of t.

Furthermore, under the conditions above the functions A and B in (13) are determined by the following system of ODEs, where the subscript t denotes partial derivative w.r.t. t, and where D denotes the block matrix $D = [D_1, \ldots, D_n]$.

$$B_t(t,T) = -B(t,T)\Delta(t) + \frac{1}{2}B(t,T)D(t)B^\star(t,T) - d(t), \tag{14}$$

$$B(T,T) = 0. \tag{15}$$

$$A_t(t,T) = B(t,T)\alpha(t,T) - \frac{1}{2}B(t,T)C(t)B^\star(t,T) + c(t), \tag{16}$$

$$A(T,T) = 0. \tag{17}$$

Proof. The proof is surprisingly simple. Given the *Ansatz* (13), and the sufficient conditions above, compute the partial derivatives and plug them into the term structure equation. The PDE will then be separable in x and the ODEs are obtained by identifying coefficients. ∎

We note that, for every fixed T, (14)-(17) is a coupled system of ODEs in the t-variable. We also see that (14) is a Riccatti equation for B, whereas (16)-(17) can be integrated directly, once B is computed.

5.2 Short rate models

The simplest type of a factor model is the one where the factor process X is scalar and coincides with the short rate, i.e., $X_t = r_t$ and $h(x) = x$. Such a model will then have the form

$$dr_t = \mu(t, r_t)dt + \sigma(t, r_t)dW_t,$$

where W is Q-Wiener. As we saw in the previous section, the term structure (i.e., the family of bond price processes) will, together with all other derivatives, be completely determined by the term structure equation

$$\frac{\partial F}{\partial t}(t,r) + \mu(t,r)\frac{\partial F}{\partial r}(t,r) + \frac{1}{2}\sigma^2(t,r)\frac{\partial^2 F}{\partial r^2}(t,r) - rF(t,r) = 0, \tag{18}$$

$$F(T,r) = \Phi(r). \tag{19}$$

In the literature there are a large number of proposals on how to specify the Q-dynamics for r. We present a (far from complete) list of the most popular models. If a parameter is time dependent this is written out explicitly. Otherwise all parameters are constant and positive.

1. Vasiček
$$dr_t = (b - ar_t)\,dt + \sigma dW_t,$$
2. Cox–Ingersoll–Ross (CIR)
$$dr_t = a\,(b - r_t)\,dt + \sigma\sqrt{r_t}dW_t,$$
3. Dothan
$$dr_t = ar_t dt + \sigma r_t dW_t,$$
4. Black–Derman–Toy (BDT)
$$dr_t = \Theta(t)r_t dt + \sigma(t)r_t dW_t,$$
5. Ho–Lee
$$dr_t = \Theta(t)dt + \sigma dW_t,$$
6. Hull–White (extended Vasiček)
$$dr_t = \{\Theta(t) - a(t)r_t\}\,dt + \sigma(t)dW_t,$$
7. Hull-White (extended CIR)
$$dr_t = \{\Theta(t) - a(t)r\}\,dt + \sigma(t)\sqrt{r_t}dW_t.$$

We now briefly comment on the models listed above.

5.2.1 Computational tractability

Looking at the list we see that all of them, apart from the Dothan and the BDT model possess affine term structures, the implication being that Dothan and the BDT model are computationally intractable.

Apart from the purely formal AFT results, there are also good probabilistic reasons why some of the models in our list are easier to handle than others. We see that the models of Vasiček, Ho–Lee and Hull–White (extended Vasiček) all describe the short rate using a **linear** SDE. Such SDEs are easy to solve and the corresponding r-processes can be shown to be normally distributed. Since bond prices are given by expressions like
$$p(0,T) = E\left[e^{-\int_0^T r(s)ds}\right],$$
and the normal property of r is inherited by the integral $\int_0^T r(s)ds$ we see that the computation of bond prices for a model with a normally distributed short rate boils down to the easy problem of computing the expected value of a log-normal stochastic variable.

In contrast with the linear models above, let us consider the Dothan model. For the Dothan model the short rate will be log-normally distributed, which means that in order to compute bond prices we are faced with determining the distribution of an integral $\int_0^T r(s)ds$ of log-normal stochastic variables. This is analytically intractable.

5.2.2 Positive short rates

From a computational point of view there is thus a lot to be said in favor of a linear SDE describing the short rate. The price we have to pay for these models is again the Gaussian property. Since the short rate will be normally distributed this means that for every t there is a positive probability that $r(t)$ is negative, and this is unreasonable from an economic point of view. For the Dothan model on the other hand, the short rate is log-normal and thus positive with probability 1. It is also possible to show that the CIR model will produce a strictly positive short rate process.

5.2.3 Mean reversion

We also note that all the models above, except Dothan, BDT, and Ho-Lee exhibit the phenomenon of **mean reversion**, i.e., the short rate has a tendency to revert to a (possibly time dependent) mean value. For example, in the Vasiček model, the drift is negative when $r > b/a$ and the drift is positive when $r < b/a$, so the short rate will revert to the long term mean b/a. Mean revertion is a typical requirement of a short rate model (as opposed to stock price models). The reason for this is basically political/institutional: if the short rate becomes very high one expects that the government and/or the central bank will intervene to lower it.

5.3 Inverting the yield curve

We now turn to the problem of parameter estimation in the martingale models above, and a natural procedure would perhaps be to use standard statistical estimation procedures based on time series data of the underlying factor process. This procedure, however, is unfortunately completely nonsensical and the reason is as follows.

Let us for simplicity assume we have a short rate model. Now, we have chosen to model the r-process by giving the Q-dynamics, which means that all parameters are defined under the martingale measure Q. When we make observations in the real world we are however **not** observing r under the martingale measure Q, but under the objective measure P. This means that

if we apply standard statistical procedures to our observed data we will not get our Q-parameters. What we get instead is pure nonsense.

To see how we can save the situation, we begin by recalling from Result 1 that **the martingale measure is chosen by the market**. Thus, in order to obtain information about the Q-drift parameters we have to **collect price information from the market**, and the typical approach is that of **inverting the yield curve** which works as follows.

- Choose a particular short rate model involving one or several parameters. (The arguments below will in fact apply to any factor model, but for simplicity we confine ourselves to short rate models). Let us denote the entire parameter vector by α. Thus we write the r-dynamics (under Q) as

$$dr_t = \mu(t, r_t; \alpha)dt + \sigma(t, r_t; \alpha)dW_t. \tag{20}$$

- Solve the term structure equation (18)-(19) to obtain the theoretical term structure as

$$p(t, T; \alpha) = F^T(t, r; \alpha).$$

- Collect price data (at $t = 0$) from the bond market for all maturities. Denote this **empirical term structure** by $\{p^\star(0, T);\ T \geq 0\}$.
- Now choose the parameter vector α in such a way that the theoretical curve $\{p(0, T; \alpha);\ T \geq 0\}$ fits the empirical curve $\{p^\star(0, T);\ T \geq 0\}$ as well as possible (according to some objective function). This gives us our estimated parameter vector $\alpha^\star$.
- We have now determined our martingale measure Q, and we can go on to compute prices of interest rate derivatives.

The procedure above is known as "inverting the yield curve", "backing out parameters from market data", or "calibrating the model to market data". It is a very general procedure, and corresponds exactly to the method of computing "implied volatilities" from option prices in a stock price model.

We end this section by noting that if we want a complete fit between the theoretical and the observed bond prices this calibration procedure is formally that of solving the system of equations

$$p(0, T; \alpha) = p^\star(0, T) \quad \text{for all } T > 0. \tag{21}$$

We observe that this is an infinite dimensional system of equations (one equation for each T) with α as the unknown, so if we work with a model containing a finite parameter vector α (like the Vasiček model) there is no hope of obtaining a perfect fit.

This is the reason why in the Hull-White model we introduce the infinite dimensional parameter vector Θ and it can in fact be shown that there exists a unique solution to (21) for the Ho-Lee model as well as for Hull-White extended Vasiček and CIR models. As an example, for the Ho-Lee model Θ is given by

$$\Theta(t) = f_T^\star(0,t) + \sigma^2 t,$$

where the lower index denotes partial derivative w.r.t maturity.

It should, however, be noted that the introduction of an infinite parameter, in order to fit the entire initial term structure, has its dangers in terms of over-parameterization, leading to numerical instability of the parameter estimates.

6 Forward Rate Models

Up to this point we have studied interest models generated by a finite number of underlying factors. The method proposed by Heath–Jarrow–Morton (HJM) is at the far end of this spectrum—they choose the entire forward rate curve as their (infinite dimensional) state variable.

6.1 The HJM drift condition

We now turn to the specification of the Heath–Jarrow–Morton framework. This can be done under P or Q, but here we confine ourselves to Q modeling.

Assumption 1 *We assume that, for every fixed $T > 0$, the forward rate $f(\cdot, T)$ has a stochastic differential which, under a given martingale measure Q, is given by*

$$df(t,T) = \alpha(t,T)dt + \sigma(t,T)dW(t), \tag{22}$$

$$f(0,T) = f^\star(0,T), \tag{23}$$

where W is a (d-dimensional) Q-Wiener process whereas $\alpha(\cdot,T)$ and $\sigma(\cdot,T)$ are adapted processes.

Note that conceptually equation (22) is a scalar stochastic differential in the t-variable for each fixed choice of T. The index T thus only serves as a "mark" or "parameter" in order to indicate which maturity we are looking at. Also note that we use the observed forward rated curve $\{f^\star(0,T);\ T \geq 0\}$ as the initial condition. This will automatically give us a perfect fit between observed and theoretical bond prices at $t = 0$, thus relieving us of the task of inverting the yield curve.

Remark 1 It is important to observe that the HJM approach to interest rates does not a propose of a specific **model**, like, for example, the Vasiček model. It is instead a **framework** to be used for analyzing interest rate models. We do not have a specific model until we have specified the drift and volatility structure in (22). Every short rate model can be equivalently

formulated in forward rate terms, and for every forward rate model, the arbitrage free price of a contingent T-claim $\mathcal{Z}$ will still be given by the pricing formula

$$\Pi(0;\mathcal{Z}) = E^Q\left[e^{-\int_0^T r(s)ds}\cdot Z\right],$$

where the short rate as usual is given by $r(s) = f(s,s)$.

We noticed earlier that for a short rate model **every** $Q \sim P$ will serve as a martingale measure. This is not the case for a forward rate model, the reason being that we have the following two different formulas for bond prices

$$p(t,T) = e^{-\int_t^T f(0,s)ds},$$
$$p(t,T) = E^Q\left[e^{-\int_0^T r(s)ds}\middle|\mathcal{F}_t\right],$$

where the short rate r and the forward rate f are connected by $r(t) = f(t,t)$. In order for these formulas to hold simultaneously, we have to impose some sort of consistency relation between α and σ in the forward rate dynamics. The result is the famous Heath–Jarrow–Morton drift condition.

Proposition 3 (HJM drift condition) *Under the martingale measure Q, the processes α and σ must satisfy the following relation, for every t and every $T \geq t$.*

$$\alpha(t,T) = \sigma(t,T)\int_t^T \sigma(t,s)^\star ds. \tag{24}$$

Proof. From Proposition 1 we obtain the bond price dynamics as

$$dp(t,T) = p(t,T)\left\{r(t) + A(t,T) + \frac{1}{2}||S(t,T)||^2\right\}dt + p(t,T)S(t,T)dW(t).$$

We also know that, under a martingale measure, the local rate of return has to equal the short rate r. Thus we obtain the identity

$$A(t,T) + \frac{1}{2}||S(t,T)||^2 = 0,$$

and differentiating this w.r.t. T gives us (24). ∎

The moral of Proposition 3 is that when we specify the forward rate dynamics (under Q) we may freely specify the volatility structure. The drift parameters are then uniquely determined.

To see at least how part of this machinery works we now study the simplest example conceivable, which occurs when the process σ is a constant. With a slight abuse of notation let us thus write $\sigma(t,T) \equiv \sigma$, where $\sigma > 0$. Equation (24) gives us the drift process as

$$\alpha(t,T) = \sigma \int_t^T \sigma ds = \sigma^2(T-t),$$

so, after integrating, equation (22) becomes

$$f(t,T) = f^\star(0,T) + \int_0^t \sigma^2(T-s)ds + \int_0^t \sigma dW(s),$$

i.e.,

$$f(t,T) = f^\star(0,T) + \sigma^2 t\left(T - \frac{t}{2}\right) + \sigma W(t).$$

In particular we see that r is given as

$$r(t) = f(t,t) = f^\star(0,t) + \sigma^2\frac{t^2}{2} + \sigma W(t),$$

so the short rate dynamics are

$$dr(t) = \left\{f_T(0,t) + \sigma^2 t\right\}dt + \sigma dW(t),$$

which we recognize as the Ho–Lee model, fitted to the initial term structure.

6.2 The Musiela parameterization

In many practical applications it is more natural to use time **to** maturity, rather than time **of** maturity, to parameterize bonds and forward rates. If we denote running time by t, time of maturity by T, and time to maturity by x, then we have $x = T - t$, and in terms of x the forward rates are defined as follows.

Definition 5 For all $x \geq 0$ the forward rates $r(t,x)$ are defined by the relation

$$r(t,x) = f(t,t+x).$$

Suppose now that we have the standard HJM-type model for the forward rates under a martingale measure Q

$$df(t,T) = \alpha(t,T)dt + \sigma(t,T)dW(t). \tag{25}$$

The question is to find the Q-dynamics for $r(t,x)$, and we have the following result, known as the Musiela equation.

Proposition 4 (The Musiela equation) *Assume that the forward rate dynamics under Q are given by (25). Then*

$$dr(t,x) = \left\{ \frac{\partial}{\partial x} r(t,x) + D(t,x) \right\} dt + \sigma_0(t,x) dW(t), \tag{26}$$

where

$$\sigma_0(t,x) = \sigma(t, t+x),$$
$$D(t,x) = \sigma_0(t,x) \int_0^x \sigma_0(t,s)' ds.$$

Proof. Using a slight variation of the Itô formula we have

$$dr(t,x) = df(t,t+x) + \frac{\partial f}{\partial T}(t,t+x)dt,$$

where the differential in the term $df(t,t+x)$ only operates on the first t. We thus obtain

$$dr(t,x) = \alpha(t,t+x)dt + \sigma(t,t+x)dW(t) + \frac{\partial}{\partial x} r(t,x)dt,$$

and, using the HJM drift condition, we obtain our result. ∎

The point of the Musiela parameterization is that it highlights equation (26) as an infinite dimensional SDE. It has become an indispensable tool of modern interest rate theory.

7 Change of Numeraire

In this section we will give a very brief account of the change of numeraire technique. We will then use the results in Section 8. All the results are standard.

7.1 Generalities

Consider a financial market (not necessarily a bond market) with the usual locally risk free asset B, and a risk neutral martingale measure Q. We recall from general theory that a measure is a martingale measure only relative to some chosen numeraire asset, and we recall that the risk neutral martingale measure, with the money account B as numeraire, has the property of martingalizing all processes of the form $S(t)/B(t)$ where S is the arbitrage free price process of any traded asset.

Assumption 2 *Assume that Q is a fixed risk neutral martingale measure, and $S_0(t)$ is a strictly positive process with the property that the process $S_0(t)/B(t)$ is a Q-martingale.*

The economic interpretation of this assumption is of course that $S_0(t)$ is the arbitrage free price process of a traded asset. We now search for a measure Q^0 with the property that, for every arbitrage free price process $\Pi(t)$, the process $\Pi(t)/S_0(t)$ is a Q^0-martingale.

In order to get an idea of what Q^0 must look like, let us consider a fixed time T and a T-contract X. Assuming enough integrability we then know that the arbitrage free price of X at time $t = 0$ is given by

$$\Pi(0; X) = E^Q\left[\frac{X}{B(T)}\right]. \tag{27}$$

Assume, on the other hand, that the measure Q^0 actually exists, with a Radon-Nikodym derivative process

$$L(t) = \frac{dQ^0}{dQ}, \quad \text{on } \mathcal{F}_t.$$

Then we know that, because of the assumed Q^0-martingale property of the process $\Pi(t; X)/S_0(t)$, we have

$$\frac{\Pi(0; X)}{S_0(0)} = E^0\left[\frac{\Pi(T; X)}{S_0(T)}\right] = E^0\left[\frac{X}{S_0(T)}\right] = E^Q\left[L(T)\frac{X}{S_0(T)}\right],$$

where E^0 denotes expectation under Q^0. Thus we have

$$\Pi(0; X) = E^Q\left[L(T)\frac{X \cdot S_0(0)}{S_0(T)}\right], \tag{28}$$

and, comparing (27) with (28), we see that a natural candidate as likelihood process for the intended change of measure is given by $L(t) = S_0(t)/S_0(0) \cdot B(t)$.

We now go on to the formal definitions and results.

Definition 6 Under Assumption 2 define, for any fixed t, the measure Q^0 on $\mathcal{F}_t$ by

$$\frac{dQ^0}{dQ} = L(t),$$

where the likelihood process L is defined by

$$L(t) = \frac{S_0(t)}{S_0(0) \cdot B(t)}. \tag{29}$$

We note at once that L is a positive Q-martingale with $L(0) = 1$, so the measure Q^0 is indeed a probability measure. We now want to prove that Q^0 martingalizes every process of the form $\Pi(t)/S_0(t)$, where $\Pi(t)$ is any arbitrage free price process. The formalization of this idea is the following result.

Proposition 5 *Define Q^0 as above. Assume that $\Pi(t)$ is a process such that $\Pi(t)/B(t)$ is a Q-martingale. Then the process $\Pi(t)/S_0(t)$ is a Q^0-martingale.*

Proof. From Bayes's formula we obtain

$$E^0\left[\frac{\Pi(t)}{S_0(t)}\middle|\mathcal{F}_s\right] = \frac{E^Q\left[L(t)\frac{\Pi(t)}{S_0(t)}\middle|\mathcal{F}_s\right]}{L(s)} = \frac{E^Q\left[\frac{\Pi(t)}{B(t)S_0(0)}\middle|\mathcal{F}_s\right]}{L(s)}$$
$$= \frac{\Pi(s)}{B(s)S_0(0)L(s)} = \frac{\Pi(s)}{S_0(s)}.$$

∎

As an immediate corollary we have the following.

Proposition 6 *Define Q^0 as above and consider a T-claim X such that $X/B(T) \in L^1(Q)$. Then the price process, $\Pi(t;X)$ is given by*

$$\Pi(t;X) = S_0(t)E^0\left[\frac{X}{S_0(T)}\middle|\mathcal{F}_t\right]. \tag{30}$$

Remark 2 Note that it is easy to find the Girsanov transformation which carries Q into Q^0. Since Q martingalizes the process $S_0(t)/B(t)$, the Q-dynamics of S_0 must be of the form

$$dS_0(t) = r(t)S_0(t)dt + S_0(t)v_0(t)dW(t), \tag{31}$$

where W is Q-Wiener, and v_0 is the volatility for S_0. From (31) and (29) it now follows that the likelihood process L has the Q-dynamics

$$dL(t) = L(t)v_0(t)dW(t), \tag{32}$$

so the relevant Girsanov kernel v_0 in (32) is in fact given by the volatility of the S_0-process.

7.2 Forward measures

In this section we specialize the theory developed in the previous section to the case when the new numeraire chosen is a bond maturing at time T. As can be expected this choice of numeraire is particularly useful when dealing with interest rate derivatives.

Suppose therefore that we are given a specified bond market model with a fixed martingale measure Q. For a fixed time of maturity T we now choose the process $p(t,T)$ as our new numeraire.

Definition 7 The **T-forward measure** Q^T is defined by

$$dQ^T = L^T(t)dQ$$

on $\mathcal{F}_t$ for $0 \leq t \leq T$ where

$$L^T(t) = \frac{p(t,T)}{B(t)p(0,T)}.$$

Observing that $P(T,T) = 1$ we have the following useful pricing formula as an immediate corollary of Proposition 6.

Proposition 7 *Assume that the T-claim X has the property that $X/B(T) \in L^1(Q)$. Then*

$$\Pi\left(t;X\right) = p(t,T)E^T\left[X \middle| \mathcal{F}_t\right],$$

where E^T denotes integration w.r.t. Q^T.

7.3 Option pricing

We will now apply the theory developed above to give a fairly general formula for the pricing of European call options. Assume therefore that we are given a financial market with a (possibly stochastic) short rate of interest r, and a strictly positive asset price process $S(t)$. We also assume the existence of a risk neutral martingale measure Q.

Consider now a fixed time T, and a European call on S with date of maturity T and strike price K. We are thus considering the T-claim

$$X = \max\left[S(T) - K,\ 0\right], \tag{33}$$

and to simplify notation we restrict ourselves to computing the price $\Pi\left(t;X\right)$ at time $t = 0$. The main trick when dealing with options is to write X as

$$X = \left[S(T) - K\right] \cdot I\left\{S(T) \geq K\right\}.$$

We obtain

$$\begin{aligned}\Pi(0;X) &= E^Q\left[B^{-1}(T)\left[S(T)-K\right]I\left\{S(T)\geq K\right\}\right]\\ &= E^Q\left[B^{-1}(T)S(T)\cdot I\left\{S(T)\geq K\right\}\right]\\ &- KE^Q\left[B^{-1}(T)\cdot I\left\{S(T)\geq K\right\}\right].\end{aligned}$$

For the first term we change to the measure Q^S having S as numeraire, and for the second term we use the T-forward measure. Using Propositions 6 and 7 we obtain the following basic option pricing formula, where we recognize the structure of the standard Black-Scholes formula.

Proposition 8 *Given the assumptions above, the option price is given by*

$$\Pi(0;X) = S(0)Q^S\left(S(T)\geq K\right) - Kp(0,T)Q^T\left(S(T)\geq K\right). \qquad (34)$$

In order to get more concrete results we make an additional assumption.

Assumption 3 *Assume that*

1. *The filtration is generated by a d-dimensional Q-Wiener process W.*
2. *The process $Z_{S,T}$ defined by*

$$Z_{S,T}(t) = \frac{S(t)}{p(t,T)},$$

has a stochastic differential of the form

$$dZ_{S,T}(t) = Z_{S,T}(t)m_T^S(t)dt + Z_{S,T}(t)\sigma_{S,T}(t)dW,$$

where the volatility process $\sigma_{S,T}(t)$ is **deterministic**.

The crucial point here is of course the assumption that the d-dimensional row vector process $\sigma_{S,T}$ is deterministic. Also note that the volatility process is unaffected by a continuous change of measure.

In order to analyze the option formula (34) we start with the second term which we write as

$$Q^T\left(S(T)\geq K\right) = Q^T\left(\frac{S(T)}{p(T,T)}\geq K\right) = Q^T\left(Z_{S,T}(T)\geq K\right).$$

By construction we know that $Z_{S,T}$ is a martingale under Q^T, so its Q^T-dynamics are given by

$$dZ_{S,T}(t) = Z_{S,T}(t)\sigma_{S,T}(t)dW^T,$$

with the solution

$$Z_{S,T}(T) = \frac{S(0)}{p(0,T)}\exp\left\{-\frac{1}{2}\int_0^T\sigma_{S,T}^2(t)dt + \int_0^T\sigma_{S,T}(t)dW^T\right\}.$$

The stochastic integral in the exponent is Gaussian with zero mean and variance

$$\Sigma_{S,T}^2(T) = \int_0^T \|\sigma_{S,T}(t)\|^2 dt. \tag{35}$$

We thus have, for the second term in (34),

$$Q^T\left(S(T) \geq K\right) = N[d_2],$$

where N denotes the cumulative distribution function of a standard Gaussian random variable, and

$$d_2 = \frac{\ln\left(\frac{S(0)}{Kp(0,T)}\right) - \frac{1}{2}\Sigma_{S,T}^2(T)}{\sqrt{\Sigma_{S,T}^2(T)}}. \tag{36}$$

For the first term in (34) we write

$$Q^S\left(S(T) \geq K\right) = Q^S\left(\frac{p(T,T)}{S(T)} \leq \frac{1}{K}\right) = Q^S\left(Y_{S,T}(T) \leq \frac{1}{K}\right),$$

where the process $Y_{S,T}$ is defined by

$$Y_{S,T}(t) = \frac{p(t,T)}{S(t)} = \frac{1}{Z_{S,T}(t)}.$$

Under the measure Q^S the process $Y_{S,T}$ is a martingale, so its Q^S-dynamics are of the form

$$dY_{S,T}(t) = Y_{S,T}(t)\delta_{S,T}(t)dW^S.$$

Since $Y_{S,T} = Z_{S,T}^{-1}$ it is easily seen that in fact $\delta_{S,T}(t) = -\sigma_{S,T}(t)$. Thus we have

$$Y_{S,T}(T) = \frac{p(0,T)}{S(0)} \exp\left\{-\frac{1}{2}\int_0^T \sigma_{S,T}^2(t)dt - \int_0^T \sigma_{S,T}(t)dW^S\right\},$$

and with exactly the same reasoning as above we have, after some simplifications,

$$Q^S\left(S(T) \geq K\right) = N[d_1],$$

where

$$d_1 = d_2 + \sqrt{\Sigma_{S,T}^2(T)}. \tag{37}$$

We have thus proved the following result.

Proposition 9 *Under the conditions given in Assumption 3, the price of the call option defined in (33) is given by the formula*

$$\Pi\left(0;X\right) = S(0)N[d_1] - K \cdot p(0,T)N[d_2], \tag{38}$$

where d_2 and d_1 are given in (36) and (37) respectively, whereas $\Sigma^2_{S,T}(T)$ is given by (35).

8 LIBOR Market Models

In the previous chapters we have concentrated on studying interest rate models based on *infinitesimal* interest rates like the instantaneous short rate and the instantaneous forward rates. These models do however suffer from some disadvantages. Firstly, the instantaneous short and forward rates can never be observed in real life. Secondly, if you would like to calibrate your model to cap or swaption data, then this is typically very complicated from a numerical point of view if you use one of the "instantaneous" models. Another disturbing fact is that, for a very long time, the market practice has been to quote caps, floors, and swaptions by using a formal extension of the Black-76 formula (see below for details). Such an extension is typically obtained by an approximation argument where the short rate at one point in the argument is assumed to be deterministic, while later on in the argument the LIBOR rate is assumed to be stochastic. This is of course logically inconsistent, but despite this, the market happily continues to use Black-76 for the pricing of caps, floors, and swaptions.

Thus there has appeared a natural demand for constructing logically consistent (and arbitrage free!) models having the property that the theoretical prices for caps, floors and swaptions produced by the model are of the Black-76 form. This project has in fact been carried out very successfully, starting with Brace et al. (1997), Jamshidian (1997), Miltersen et al. (1997). The basic structure of the models is as follows.

Instead of modeling instantaneous interest rates, we model discrete **market rates** like LIBOR rates. Under a suitable choice of numeraire(s), these market rates can in fact be modeled log normally. The market models will thus produce pricing formulas for caps and floors (the LIBOR models) which are of the Black-76 type and thus conforming with market practice. By construction the market models are thus very easy to calibrate to market data for caps/floors and swaptions respectively. They are then used to price more exotic products. For this later pricing part, however, we will typically have to resort to some numerical method, like Monte Carlo.

8.1 Caps: definition and market practice

In this section we discuss LIBOR caps and the market practice for pricing and quoting these instruments. To this end we consider a fixed set of increasing maturities $T_0, T_1, \ldots, T_N$ and we define α_i, by

$$\alpha_i = T_i - T_{i-1}, \quad i = 1, \ldots, N.$$

The number α_i is known as the **tenor**, and in a typical application we could for example have all α_i equal to a quarter of a year.

Definition 8 We let $p_i(t)$ denote the zero coupon bond price $p(t, T_i)$ and let $L_i(t)$ denote the LIBOR forward rate, contracted at t, for the period $[T_{i-1}, T_i]$, i.e.,

$$L_i(t) = \frac{1}{\alpha_i} \cdot \frac{p_{i-1}(t) - p_i(t)}{p_i(t)}, \quad i = 1, \ldots, N.$$

We recall that a **cap** with **cap rate** R and **resettlement dates** $T_0, \ldots, T_N$ is a contract which at time T_i gives the holder of the cap the amount

$$X_i = \alpha_i \cdot \max\left[L_i(T_{i-1}) - R, 0\right],$$

for each $i = 1, \ldots, N$. The cap is thus a portfolio of the individual **caplets** $X_1, \ldots, X_N$. We note that the forward rate $L_i(T_{i-1})$ above is in fact the spot rate at time T_{i-1} for the period $[T_{i-1}, T_i]$, and is determined already at time T_{i-1}. The amount X_i is thus determined at T_{i-1} but not payed out until at time T_i. We also note that, formally speaking, the caplet X_i is a call option on the underlying spot rate.

The market practice is to use the Black-76 formula for the pricing of caplets.

Definition 9 (Black's formula for caplets)
The Black-76 formula for the caplet

$$X_i = \alpha_i \cdot \max\left[L(T_{i-1}, T_i) - R, 0\right],$$

is given by the expression

$$\mathbf{Capl_i^B}(t) = \alpha_i \cdot p_i(t) \left\{L_i(t) N[d_1] - R N[d_2]\right\}, \quad i = 1, \ldots, N,$$

where

$$d_1 = \frac{1}{\sigma_i \sqrt{T_i - t}} \left[\ln\left(\frac{L_i(t)}{R}\right) + \frac{1}{2}\sigma_i^2 (T - t)\right],$$
$$d_2 = d_1 - \sigma_i \sqrt{T_i - t}.$$

The constant σ_i is known as the **Black volatility** for caplet No. i. In order to make the dependence on the Black volatility σ_i explicit we will sometimes write the caplet price as $Capl_i^B(t; \sigma_i)$.

In the market, cap prices are not quoted in monetary terms but instead in terms of **implied Black volatilities**, and these volatilities can furthermore be quoted as **flat volatilities** or as **spot volatilities**. Here, we confine ourselves to spot volatilities.

Consider a fixed date t, a fixed set of dates $T_0, T_1, \ldots, T_N$ where $t \leq T_0$, and a fixed cap rate R. We assume that we can observe the market prices $\mathbf{Capl_i^m}(t), \quad i = 1, \ldots, N$ for the corresponding caplets.

Definition 10 Given market price data as above, the implied Black volatilities $\bar{\sigma}_1, \ldots, \bar{\sigma}_N$ are defined as the solutions of the equations

$$\mathbf{Capl_i^m}(t) = \mathbf{Capl_i^B}(t; \bar{\sigma}_i), \quad i = 1, \ldots, N.$$

A sequence of implied volatilities $\bar{\sigma}_1, \ldots, \bar{\sigma}_N$ is called a volatility **term structure**.

8.2 The LIBOR market model

We now turn from market practice to the construction of the so-called LIBOR market models. To motivate these models let us consider the theoretical arbitrage free pricing of caps. Using the T_i forward measure Q^{T_i} (for short Q^i), the price $c_i(t)$ of a caplet No. i is

$$\mathbf{Capl_i}(t) = \alpha_i p_i(t) E^{T_i}\left[\max\left[L_i(T_{i-1}) - R, 0\right] \middle| \mathcal{F}_t\right], \quad i = 1, \ldots, N. \tag{39}$$

The focal point of the LIBOR models is the following simple result.

Lemma 3 *For every $i = 1, \ldots, N$, the LIBOR process L_i is a martingale under the corresponding forward measure Q^{T_i}, on the interval $[0, T_{i-1}]$.*

Proof. We have

$$\alpha_i \cdot L_i(t) = \frac{p_{i-1}(t)}{p_i(t)} - 1.$$

The process 1 is obviously a martingale under any measure. The process p_{i-1}/p_i is the price of the T_{i-1} bond normalized by the numeraire p_i. Since p_i is the numeraire for the martingale measure Q^{T_i}, the process p_{i-1}/p_i is thus trivially a martingale on the interval $[0, T_{i-1}]$. Thus $\alpha_i L_i$ is a martingale and hence L_i is also a martingale. ∎

The basic idea is now to define the LIBOR rates such that, for each i, $L_i(T)$ will be lognormal under "its own" measure Q^i, since then all caplet prices in (39) will be given by a Black type formula. The formal definition is as follows.

Definition 11 If the LIBOR forward rates have the dynamics

$$dL_i(t) = L_i(t)\sigma_i(t)dW^i(t), \quad i = 1, \ldots, N, \tag{40}$$

where W^i is Q^i-Wiener, and σ_i is deterministic, then we say that we have a discrete tenor **LIBOR market model**.

8.3 Pricing caps in the LIBOR model

Given a LIBOR market model, the pricing of a caplet, and hence also a cap, is trivial. Since L_i in (40) is just a geometrical Brownian motion (GBM) we obtain

$$L_i(T) = L_i(t) \cdot e^{\int_t^T \sigma_i(s)dW^i(s) - \frac{1}{2}\int_t^T \|\sigma_i(s)\|^2 ds}.$$

Since σ_i is assumed to be deterministic this implies that, conditional on $\mathcal{F}_t$, $L_i(T)$ is lognormal, and a simple calculation gives us the following pricing formula for caps.

Proposition 10 *In the LIBOR market model, the caplet prices are given by*

$$\mathbf{Capl_i}(t) = \alpha_i \cdot p_i(t) \left\{ L_i(t)N[d_1] - RN[d_2] \right\}, \quad i = 1, \ldots, N,$$

where

$$\begin{aligned} d_1 &= \frac{1}{\Sigma_i(t, T_{i-1})} \left[\ln\left(\frac{L_i(t)}{R}\right) + \frac{1}{2}\Sigma_i^2(t, T_{i-1}) \right], \\ d_2 &= d_1 - \Sigma_i(t, T_{i-1}), \end{aligned}$$

with Σ_i defined by

$$\Sigma_i^2(t,T) = \int_t^T \|\sigma_i(s)\|^2 ds.$$

We thus see that each caplet price is given by a Black type formula.

8.4 Terminal measure dynamics and existence

We now turn to the question whether there always exists a LIBOR market model for any given specification of the deterministic volatilities $\sigma_1, \ldots, \sigma_N$. In order to get started we first have to specify all LIBOR rates $L_1, \ldots, L_N$ under **one** common measure, and the canonical choice is the **terminal measure** Q^N.

After long and tedious calculations, the following existence result can be proved.

Proposition 11 *Consider a given volatility structure σ_1, σ_N, where each σ_i is assumed to be bounded, a probability measure Q^N and a standard Q^N-Wiener process W^N. Define the processes $L_1, \ldots, L_N$ by*

$$dL_i(t) = -L_i(t) \left(\sum_{k=i+1}^{N} \frac{\alpha_k L_k(t)}{1 + \alpha_k L_k(t)} \sigma_k(t) \sigma_i^\star(t) \right) dt + L_i(t)\sigma_i(t) dW^N(t),$$

for $i = 1, \ldots, N$ where we use the convention $\sum_N^N(\ldots) = 0$. Then the Q^i-dynamics of L_i are given by (40). Thus there exists a LIBOR model with the given volatility structure.

9 Potentials and Positive Interest

The purpose of this section is to present two approaches to interest rate theory based on so called "stochastic discount factors" (see below for details), while also relating bond pricing to stochastic potential theory.

An appealing aspect of the approaches described below is that they both generate **positive term structures**, i.e., a system of bond prices for which all induced forward rates are positive.

9.1 Generalities

As a general setup we consider a standard filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, P)$ where P is the objective measure. We now need an assumption about how the market prices various assets.

Assumption 4 *We assume that the market prices all assets, underlying and derivative, using a fixed martingale measure Q (with the money account as the numeraire).*

We now recall that for a T-claim Y the arbitrage free price at $t = 0$ is given by

$$\Pi(0; Y) = E^Q\left[e^{-\int_0^T r_s ds} \cdot Y\right]. \tag{41}$$

We denote the likelihood process for the transition from the objective measure P to the martingale measure Q by L, i.e.,

$$L_t = \frac{dQ_t}{dP_t},$$

where the index t denotes the restriction of P and Q to $\mathcal{F}_t$. We may of course also write the price in (41) as an expected value under P:

$$E^P\left[e^{-\int_0^T r_s ds} \cdot L_T \cdot Y\right].$$

This leads us to the following definition.

Definition 12 The **stochastic discount factor** (SDF), or **state price density process** Z is defined by

$$Z(t) = e^{-\int_0^t r_s ds} \cdot L_t.$$

We now have the following basic pricing result, which follows directly from the Bayes formula.

Proposition 12 *For any T-claim Y, the arbitrage free price process is given by*

$$\Pi(t;X) = \frac{E^P[Z_T Y|\mathcal{F}_t]}{Z_t}.$$

In particular, bond prices are given by

$$\Pi(t;X) = \frac{E^P[Z_T|\mathcal{F}_t]}{Z_t}. \tag{42}$$

We now have the following fact which we will use extensively.

Proposition 13 *Assume that the short rate is strictly positive and that the economically natural condition $p(0,T) \to 0$ as $T \to \infty$ is satisfied. Then the stochastic discount factor Z is a probabilistic* **potential**, *i.e.,*

- *Z is a supermartingale.*
- *$E[Z_t] \to 0$ as $t \to \infty$.*

Conversely one can show that any potential will serve as a stochastic discount factor. Thus the moral is that modeling bond prices in a market with positive interest rates is equivalent to modeling a potential, and in the next sections we will describe two ways of doing this.

We end by noticing that we can easily recover the short rate from the dynamics of Z.

Proposition 14 *If the dynamics of Z are written as*

$$dZ_t = -h_t dt + dM_t,$$

where h is nonnegative and M is a martingale, then the short rate is given by

$$r_t = Z_t^{-1} h_t.$$

Proof. Applying the Itô formula to the definition of Z we obtain

$$dZ_t = -r_t Z_t dt + e^{-\int_0^t r_s ds} dL_t. \quad \blacksquare$$

9.2 The Flesaker–Hughston fractional model

Given a stochastic discount factor Z and a positive short rate we may, for each fixed T, define the process $\{X(t,T);\ 0 \leq t \leq T\}$ by

$$X(t,T) = E^P\left[Z_T \middle| \mathcal{F}_t\right], \tag{43}$$

and thus, according to (42) write bond prices as

$$p(t,T) = \frac{X(t,T)}{X(t,t)}. \tag{44}$$

We now have the following result.

Proposition 15 *For each fixed t, the mapping $T \longmapsto X(t,T)$ is smooth, and in fact*

$$\frac{\partial}{\partial T}X(t,T) = -E^P\left[r_T Z_T \middle| \mathcal{F}_t\right]. \tag{45}$$

Furthermore, for each fixed T, the process

$$X_T(t,T) = \frac{\partial}{\partial T}X(t,T)$$

is a negative P-martingale satisfying

$$X_T(0,T) = -p_T(0,T), \quad \textit{for all } T \geq 0.$$

Proof. Using the definition of Z and the Itô formula, we obtain

$$dZ_s = -r_s Z_s ds + Z_s dL_s,$$

so

$$Z_T = Z_t - \int_t^T r_s Z_s ds + \int_t^T Z_s dL_s.$$

Since L is a martingale, this gives us

$$E^P\left[Z_T \middle| \mathcal{F}_t\right] = -E^P\left[\int_t^T r_s Z_s ds \middle| \mathcal{F}_t\right],$$

and (45) follows immediately. The martingale property now follows directly from (45). ∎

We can now state the basic result from Flesaker–Hughston.

Theorem 7 *Assume that the term structure is positive. Then there exists a family of positive martingales $M(t,T)$ indexed by T and a positive deterministic function Φ such that*

$$p(t,T) = \frac{\int_T^\infty \Phi(s)M(t,s)ds}{\int_t^\infty \Phi(s)M(t,s)ds}. \tag{46}$$

The M family can, up to multiplicative scaling by the Φ process, be chosen as

$$M(t,T) = -X_T(t,T) = E^P\left[r_T Z_T \middle| \mathcal{F}_t\right].$$

In particular, Φ can be chosen as

$$\Phi(s) = -p_T(0,s), \tag{47}$$

in which case the corresponding M is normalized to $M(0,s) = 1$ for all $s \geq 0$.

Proof. A positive term structure implies that $X(t,T) \to 0$ as $T \to \infty$, so we have

$$X(t,T) = -\int_T^\infty X_T(t,s)ds,$$

and thus we obtain from (44)

$$p(t,T) = \frac{\int_T^\infty X_T(t,s)ds}{\int_t^\infty X_T(t,s)ds}. \tag{48}$$

If we now define $M(t,T)$ by

$$M(t,T) = -X_T(t,T),$$

then (46) follows from (48) with $\Phi \equiv 1$. The function Φ is only a scale factor which can be chosen arbitrarily, and the choice in (47) is natural in order to normalize the M family. Since X_T is negative, M is positive and we are done. ∎

There is also a converse of the result above.

Proposition 16 *Consider a given family of positive martingales $M(t,T)$ indexed by T and a positive deterministic function Φ. Then the specification*

$$p(t,T) = \frac{\int_T^\infty \Phi(s)M(t,s)ds}{\int_t^\infty \Phi(s)M(t,s)ds}, \tag{49}$$

defines an arbitrage free positive system of bond prices. Furthermore, the stochastic discount factor Z generating the bond prices is given by

$$Z_t = \int_t^\infty \Phi(s)M(t,s)ds.$$

Proof. Using the martingale property of the M family, we obtain

$$E^P\left[Z_T|\,\mathcal{F}_t\right] = \int_T^\infty E^P\left[\Phi(s)M(T,s)|\,\mathcal{F}_t\right]ds = \int_T^\infty \Phi(s)M(t,s)ds.$$

This implies, by the positivity of M and Φ, that Z is a potential and can thus serve as a stochastic discount factor. The induced bond prices are thus given by

$$p(t,T) = \frac{E^P\left[Z_T|\,\mathcal{F}_t\right]}{Z_t},$$

and the calculation above shows that the induced (arbitrage free) bond prices are given by (49). ∎

The most used instance of a Flesaker-Hughston model is the so called **rational model**. In such a model we consider a given martingale K and two deterministic positive functions $\alpha(t)$ and $\beta(t)$. We then define the M family by

$$M(t,T) = \alpha(T) + \beta(T)K(t).$$

With this specification of M it is easily seen that bond prices will have the form

$$p(t,T) = \frac{A(T) + B(T)K(t)}{A(t) + B(t)K(t)},$$

where

$$A(t) = \int_t^\infty \Phi(s)\alpha(s)ds, \quad B(t) = \int_t^\infty \Phi(s)\beta(s)ds,$$

We can specialize this further by assuming K to be of the form

$$K(t) = e^{\int_0^t \gamma(s)dW_s - \frac{1}{2}\int_0^t \gamma^2(s)ds},$$

where γ is deterministic. Then K will be a lognormal martingale, and the entire term structure will be analytically very tractable.

9.3 Connections to the Riesz decomposition

In Section 9.1 we saw that any stochastic discount factor generating a nice bond market is a potential, so from a modeling point of view it is natural to ask how one can construct potentials from scratch.

The main tool used is the following standard result.

Proposition 17 (Riesz Decomposition) *If Z is a potential, then it admits a representation as*

$$Z_t = -A_t + M_t, \tag{50}$$

where A is an increasing process, and M is a martingale defined by

$$M_t = E^P\left[A_\infty|\,\mathcal{F}_t\right].$$

To construct a potential, let us assume that we define A as

$$A_t = \int_0^t a_s ds \tag{51}$$

for some integrable nonnegative process a. Then we easily obtain

$$Z_t = E^P\left[\int_0^\infty a_s ds \middle| \mathcal{F}_t\right] - \int_0^t a_s ds = \int_t^\infty E^P\left[a_s \middle| \mathcal{F}_t\right] ds. \tag{52}$$

We can now connect this to the Flesaker-Hughston framework. The family of processes $X(t,T)$ defined in (43) will, in the present framework, have the form

$$X(t,T) = E^P\left[\int_T^\infty E^P\left[a_s \middle| \mathcal{F}_T\right] ds \middle| \mathcal{F}_t\right] = \int_T^\infty E^P\left[a_s \middle| \mathcal{F}_t\right] ds,$$

so the basic family of Flesaker-Hughston martingales are given by

$$M(t,T) = -\frac{\partial}{\partial T} X(t,T) = E^P\left[a_T \middle| \mathcal{F}_t\right].$$

9.4 Conditional variance potentials

An alternative way of representing potentials which has been studied in depth by Hughston and co-authors is through conditional variances.

Consider a fixed random variable $X_\infty \in L^2(P, \mathcal{F}_\infty)$. We can then define a martingale X by setting

$$X_t = E^P\left[X_\infty \middle| \mathcal{F}_t\right].$$

Now let us define the process Z by

$$Z_t = E^P\left[\left(X_\infty - X_t\right)^2 \middle| \mathcal{F}_t\right].$$

An easy calculation shows that

$$Z_t = E^P\left[X_\infty^2 \middle| \mathcal{F}_t\right] - X_t^2.$$

Since the first term is a martingale and the second is a submartingale, the difference is a supermartingale, which by definition is positive and it is in fact a potential.

The point of this is that the potential Z, and thus the complete interest rate model generated by Z, is in fact fully specified by a specification of the

single random variable X_∞. A very interesting idea is now to expand X_∞ into Wiener chaos. See the notes in Section 10 below.

9.5 The Rogers Markov potential approach

As we have seen above, in order to generate an arbitrage free bond market model it is enough to construct a positive supermartingale which acts as stochastic discount factor, and in the previous section we saw how to do this using the Riesz decomposition. In this section we will present a systematic way of constructing potentials along the lines above, in terms of Markov processes and their resolvents. The ideas are due to Rogers, and we largely follow his presentation.

We consider a time homogeneous Markov process X under the objective measure P, with infinitesimal generator $\mathcal{G}$.

For any positive real valued sufficiently integrable function g and any positive number α we can now define the process A in the Riesz decomposition (50) as

$$A_t = \int_0^t e^{-\alpha s} g(X_s) ds,$$

where the exponential is introduced in order to allow for at least all bounded functions g. In terms of the representation (51) we thus have

$$a_t = e^{-\alpha t} g(X_t),$$

and a potential Z is, according to (52), obtained by

$$Z_t = \int_t^\infty e^{-\alpha s} E^P \left[g(X_s) | \, \mathcal{F}_t \right] ds.$$

Using the Markov assumption we thus have

$$Z_t = E^P \left[\int_t^\infty e^{-\alpha s} g(X_s) ds \middle| \, X_t \right], \tag{53}$$

and this expression leads to a well known probabilistic object.

Definition 13 For any nonnegative α the **resolvent** R_α is an operator, defined for any bounded measurable function g by the expression

$$R_\alpha g(x) = E_x^P \left[\int_0^\infty e^{-\alpha s} g(X_s) ds \right],$$

where subscript x refers to the conditioning $X_0 = x$.

We can now connect resolvents to potentials.

Proposition 18 *For any bounded nonnegative g, the process*

$$Z_t = e^{-\alpha t}\frac{R_\alpha g(X_t)}{R_\alpha g(X_0)} \tag{54}$$

is a potential with $Z_0 = 1$.

Proof. The normalizing factor is trivial so we disregard it in the rest of the proof. Using time invariance we have, from (53),

$$Z_t = E^P\left[\int_0^\infty e^{-\alpha(t+s)}g(X_{t+s})ds\middle| X_t\right] = e^{-\alpha t}R_\alpha g(X_t).$$

∎

Given a SDF of the form above, we can of course compute bond prices, and the short rate can easily be recovered.

Proposition 19 *If the stochastic discount factor Z is defined by (54) then bond prices are given by*

$$p(t,T) = e^{-\alpha(T-t)}\frac{E^P\left[R_\alpha g(X_T)|\,\mathcal{F}_t\right]}{R_\alpha g(X_t)}, \tag{55}$$

and the short rate is given by

$$r_t = \frac{g(X_t)}{R_\alpha g(X_t)}. \tag{56}$$

Proof. The formula (55) follows directly from the general formula (42). From the construction of the process a we have

$$dZ_t = -e^{-\alpha t}g(X_t)dt + dM_t,$$

and (56) now follows from Proposition 14. ∎

One problem with this scheme is that, for a concrete case, it may be very hard to compute the quotient in (56). To overcome this difficulty we recall the following standard result.

Proposition 20 *With notation as above we have essentially*

$$R_\alpha = (\alpha - \mathcal{G})^{-1}. \tag{57}$$

The phrase "essentially" indicates that the result is "morally" correct, but that care has to be taken concerning the domain of the operators.

Using the identity $R_\alpha = (\alpha - \mathcal{G})^{-1}$ we see that with $f = R_\alpha g$ we have

$$\frac{g(X_t)}{R_\alpha g(X_t)} = \frac{(\alpha - \mathcal{G})f(X_t)}{f(X_t)},$$

where it usually is a trivial task to compute the last quotient.

This led Rogers to use the following scheme.

1. Fix a Markov process X, number α and a nonnegative function f.
2. Define g by
$$g = (\alpha - \mathcal{G})\, f.$$
3. Choose α (and perhaps the parameters of f) such that g is nonnegative.
4. Now we have $f = R_\alpha g$, and the short rate can be recaptured by
$$r(t) = \frac{(\alpha - \mathcal{G})f(X_t)}{f(X_t)}.$$

In this way Rogers produces a surprising variety of concrete analytically tractable nonnegative interest rate models and, exchange rate models can also be treated within the same framework.

10 Notes

All basic material in this article can be found in most advanced textbooks, like Björk (2004) and Duffie (2001). The martingale approach to arbitrage pricing was developed in Harrison and Kreps (1979) and Harrison and Pliska (1981). It was then extended in, among other papers, Delbaen and Schachermayer (1994), Duffie and Huang (1986). An elementary textbook on bond markets is Fabozzi (2004). For more advanced treatments see Björk (2004) and Duffie (2001). The encyclopedic book Brigo and Mercurio (2001) contains a wealth of theoretical, numerical and and practical information. Basic papers on short rate models are Cox et al. (1985), Ho and Lee (1986), Hull and White (1990), Vasiček (1977). For an example of a two-factor model see Longstaff and Schwartz (1992). For extensions and notes on the affine term structure theory, see Duffie and Kan (1996). Jump processes (and affine theory) is treated in Björk (1997), Duffie et al. (2000). The HJM framework first appeared in Heath et al. (1992) and the Musiela parameterization first appeared in Brace and Musiela (1994). The change of numeraire was introduced in Margrabe (1978), and developed more systematically in Geman et al. (1995), Jamshidian (1989). The LIBOR market models were developed in Brace et al. (1997), Miltersen et al. (1997). See also Jamshidian (1997) for swap market models. The Flesaker-Hughston models appeared in Flesaker and Hughston (1996) and analyzed further in Jin and Glasserman (2001). For the Rogers potential approach, see Rogers (1994).

References

Björk, T. (2004): *Arbitrage Theory in Continuous Time*, 2nd ed. Oxford University Press.

Björk, T., Kabanov, Y. and Runggaldier, W. (1997): Bond market structure in the presence of a marked point process. *Mathematical Finance* **7**, 211–239.

Brace, A., Gatarek, D. and Musiela, M. (1997): The market model of interest rate dynamics. *Mathematical Finance* **7**, 127–154.

Brace, A. and Musiela, M. (1994): A multifactor Gauss Markov implementation of Heath, Jarrow, and Morton. *Mathematical Finance* **4**, 259–283.

Brigo, D. and Mercurio, F. (2001): *Interest Rate Models*. Springer.

Cox, J., Ingersoll, J. and Ross, S. (1985): A theory of the term structure of interest rates. *Econometrica* **53**, 385–407.

Delbaen, F. and Schachermayer, W. (1994): A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* **300**, 215–250.

Duffie, D. (2001): *Dynamic Asset Pricing Theory*, 3rd ed. Princeton University Press.

Duffie, D. and Huang, C. (1986): Multiperiod securities markets with differential information. *Journal of Mathematical Economics* **15**, 283–303.

Duffie, D. and Kan, R. (1996): A yield-factor model of interest rates. *Mathematical Finance* **6**, 379–406.

Duffie, D., Pan, J. and Singleton, K. (2000): Transform analysis and asset pricing for affine jump diffusions. *Econometrica* **68**, 1343–1376.

Fabozzi, F. (2004): *Bond markets, Analysis, and Strategies*. Prentice Hall.

Flesaker, B. and Hughston, L. (1996): Positive interest. *RISK Magazine* **9**, 46–49.

Geman, H., El Karoui, N. and Rochet, J.-C. (1995): Changes of numéraire, changes of probability measure and option pricing. *Journal of Applied Probability* **32**, 443–458.

Harrison, J. and Kreps, J. (1979): Martingales and arbitrage in multiperiod markets. *Journal of Economic Theory* **11**, 418–443.

Harrison, J. and Pliska, S. (1981): Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications* **11**, 215–260.

Heath, D., Jarrow, R. and Morton, A. (1992): Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation. *Econometrica* **60**, 77–105.

Ho, T. and Lee, S. (1986): Term structure movements and pricing interest rate contingent claims. *Journal of Finance* **41**, 1011–1029.

Hull, J. and White, A. (1990): Pricing interest-rate-derivative securities. *Review of Financial Studies* **3**, 573–592.

Jamshidian, F. (1989): An exact bond option formula. *Journal of Finance* **44**, 205–209.

Jamshidian, F. (1997): LIBOR and swap market models and measures. *Finance and Stochastics* **1**, 293–330.

Jin, Y. and Glasserman, P. (2001): Equilibrium positive interest rates: a unified view. *Review of Financial Studies* **14**, 187–214.

Longstaff, F. and Schwartz, E. (1992): Interest rate volatility and the term structure. *Journal of Finance* **40**, 1259–1282.

Margrabe, W. (1978): The value of an option to exchange one asset for another. *Journal of Finance* **33**, 177–186.

Miltersen, K., Sandmann, K. and Sondermann, D. (1997): Closed form solutions for term structure derivatives with log-normal interest rates. *Journal of Finance* **52**, 409–430.

Rogers, L.C.G. (1994): The potential approach to the term structure of interest rates and foreign exchange rates. *Mathematical Finance* **7**, 157–176.

Vasiček, O. (1977): An equilibrium characterization of the term stucture. *Journal of Financial Economics* **5**, 177–188.

Extremes of Continuous–Time Processes

Vicky Fasen *

Abstract In this paper we present a review on the extremal behavior of stationary continuous-time processes with emphasis on generalized Ornstein-Uhlenbeck processes. We restrict our attention to heavy-tailed models like heavy-tailed Ornstein-Uhlenbeck processes or continuous-time GARCH processes. The survey includes the tail behavior of the stationary distribution, the tail behavior of the sample maximum and the asymptotic behavior of sample maxima of our models.

1 Introduction

In this paper we study the extremal behavior of stationary continuous-time processes. The class of stationary continuous-time processes is rich, and the investigation of their extremal behavior is complex. The development of the extremal behavior of Gaussian processes, which is the origin of continuous-time extreme value theory starting with Rice (1939, 1944, 1945), Kac (1943), Kac and Slepian (1959), Volkonskii and Rozanov (1959, 1961) and Slepian (1961, 1962), alone, would fill a paper. See the monograph of Leadbetter et al. (1983) or the Ph.D. thesis of Albin (1987) or the paper Albin (1990) for a review of this topic. Since financial time series

- are often random with jumps,
- have heavy tails,
- exhibit clusters on high levels,

Vicky Fasen
Technische Universität München, Zentrum Mathematik, Boltzmannstrasse 3, D–85747 Garching, e-mail: fasen@ma.tum.de

* Parts of the paper were written while the author was visiting the Department of Operations Research and Information Engineering at Cornell University. She takes pleasure in thanking the colleagues there for their hospitality. Financial support by the Deutsche Forschungsgemeinschaft through a research grant is gratefully acknowledged.

DOI: 10.1007/978-3-540-71297-8_28,

we will concentrate mainly on stationary continuous-time processes having these properties.

We will explain the basic ideas concerning extreme value theory for stationary continuous-time processes by generalized Ornstein-Uhlenbeck (GOU) processes, which are applied as stochastic volatility models in finance and as risk models in insurance. They are represented by

$$X_t = \mathrm{e}^{-\xi_t} \int_0^t \mathrm{e}^{\xi_{s-}}\, d\eta_s + \mathrm{e}^{-\xi_t} X_0, \quad t \geq 0, \tag{1}$$

where $(\xi_t, \eta_t)_{t\geq 0}$ is a bivariate Lévy process independent of the starting random variable X_0 (cf. Lindner and Maller (2005) and for definitions, further details and references see also Maller et al. (2008) in this volume). A bivariate Lévy process is characterized by the *Lévy-Khinchine representation*

$$\mathbb{E}(\mathrm{e}^{i\langle \Theta, (\xi_t, \eta_t)\rangle}) = \exp(-t\Psi(\Theta)) \quad \text{for } \Theta \in \mathbb{R}^2,$$

where

$$\Psi(\Theta) = -i\langle \gamma, \Theta\rangle + \frac{1}{2}\langle \Theta, \Sigma\, \Theta\rangle + \int_{\mathbb{R}^2} \Big(1 - \mathrm{e}^{i\langle \Theta,(x,y)\rangle} + i\langle (x,y), \Theta\rangle\Big)\, d\Pi_{\xi,\eta}(x,y)$$

with $\gamma \in \mathbb{R}^2$, Σ a non-negative definite matrix in $\mathbb{R}^{2\times 2}$, $\langle \cdot, \cdot \rangle$ the inner product and $\Pi_{\xi,\eta}$ a measure on $\mathbb{R}^2$, called *Lévy measure*, such that $\int_{\mathbb{R}^2} \min\{\sqrt{x^2+y^2}, 1\}\, d\Pi_{\xi,\eta}(x,y) < \infty$ and $\Pi_{\xi,\eta}((0,0)) = 0$ (cf. Sato (1999)).

The limit behavior of the sample maxima

$$M(T) = \sup_{0\leq t\leq T} X_t \tag{2}$$

as $T \to \infty$ of the stationary GOU-process $(X_t)_{t\geq 0}$ will be described either when $\xi_t = \lambda t$ or when $\mathbb{E}(\mathrm{e}^{-\alpha\xi_1}) = 1$ for some $\alpha > 0$.

In Section 2, a synopsis of extreme value theory is given. Precise definitions of the GOU-models studied in this paper are presented in Section 3. We start with the investigation of the tail behavior of the sample maximum in Section 4. Section 5 on the asymptotic behavior of sample maxima $M(T)$ as $T \to \infty$ and the cluster behavior follows. Finally, Section 6 concludes with remarks on extensions of the results to more general models.

2 Extreme Value Theory

One method of investigating extremes of stationary continuous-time processes is to study the extremal behavior of the discrete-time skeleton

$$M_k(h) = \sup_{(k-1)h \le s \le kh} X_s \quad \text{for } k \in \mathbb{N} \tag{3}$$

and some fixed $h > 0$, which is again a stationary sequence. The advantage of such a skeleton is that known results for sequences can be applied, which are well investigated; see De Haan and Ferreira (2006), Embrechts et al. (1997), Leadbetter et al. (1983) and Resnick (1987). To my knowledge this idea was first applied to Gaussian processes by Leadbetter and Rootzén (1982). The monograph of Leadbetter et al. (1983) and the paper of Leadbetter and Rootzén (1988) contain a detailed study of extremes of discrete-time and continuous-time processes. A completely different approach to extreme value theory for continuous-time processes as presented here is given in Berman (1992). Both approaches were combined by Albin (1987, 1990).

2.1 Extremes of discrete–time processes

We start with an introduction into extremes of discrete-time processes. Let $(Y_n)_{n\in\mathbb{N}}$ be a stationary sequence with distribution function F and $M_n = \max\{Y_1, \ldots, Y_n\}$ for $n \in \mathbb{N}$. The simplest stationary sequence is an iid (independently and identically distributed) sequence. In this case, we find sequences of constants $a_n > 0$, $b_n \in \mathbb{R}$, such that

$$\lim_{n\to\infty} \mathbb{P}(a_n^{-1}(M_n - b_n) \le x) = G(x) \quad \text{for } x \in \text{supp}(G), \tag{4}$$

and some non-degenerate distribution function G whose support is denoted by $\text{supp}(G)$, if and only if

$$\lim_{n\to\infty} n\overline{F}(a_n x + b_n) = -\log G(x) \quad \text{for } x \in \text{supp}(G)\,, \tag{5}$$

where $\overline{F} = 1 - F$ denotes the *tail* of F. Then we say that F is in the *maximum domain of attraction* of G ($F \in \text{MDA}(G)$). The Extremal Types Theorem (Leadbetter et al. (1983), Theorem 1.4.2) says that G is either a Fréchet (Φ_α, $\alpha > 0$), a Gumbel (Λ) or a Weibull (Ψ_α, $\alpha > 0$) distribution.

For a stationary sequence $(Y_n)_{n\in\mathbb{N}}$ there exists sufficient conditions such that the extremal behavior of the stationary sequence coincides with the extremal behavior of an iid sequence with the same stationary distribution; i.e. (5) implies (4). The conditions which guarantee this conclusion are known as D and D' conditions (cf. Leadbetter et al. (1983), pp. 53). The condition D is a mixing condition for the asymptotic independence of maxima, and the condition D' is an anti-clustering condition. That is, given an observation at some time n is large, the probability that any of the neighboring observations are also large is quite low.

Examples exist which do not satisfy the D' condition and which have extremal clusters on high level values. There, the *extremal index* is defined as

a measure of the cluster size; i. e. if (5) holds and

$$\lim_{n\to\infty} \mathbb{P}(a_n^{-1}(M_n - b_n) \le x) = G^\theta(x) \quad \text{for } x \in \text{supp}(G),$$

then θ is called the extremal index. The parameter θ takes only values in $[0, 1]$, where $\theta = 1$ reflects no extremal clusters.

2.2 Extremes of continuous–time processes

After these basic ideas concerning extremes of stationary discrete-time processes, we continue with extremes of stationary continuous-time processes. The extremal behavior of a continuous-time process is influenced by the dependence of the process not only in large, but also in small time intervals. The dependence structure of the process in small time intervals is negated by investigating the extremal behavior of $(M_k(h))_{k\in\mathbb{N}}$ as in (3), where $\max_{k=1,\dots,n} M_k(h) = M(nh)$. The conditions D and D' on the sequence $(M_k(h))_{k\in\mathbb{N}}$ can be reformulated as conditions on the continuous-time process $(X_t)_{t\ge 0}$ known as C and C' conditions. Again, condition C is a condition on the asymptotic independence of maxima and C' on the cluster behavior of $(M_k(h))_{k\in\mathbb{N}}$. Similar to discrete-time models, an Extremal Types Theorem also holds (cf. Leadbetter et al. (1983), Theorem 13.1.5). For Gaussian processes a simple condition only on the covariance function exists such that C and C' are satisfied (cf. Leadbetter et al. (1983), Theorem 12.3.4).

As in discrete time, there are also continuous-time examples which do not satisfy the C' condition and have extremal clusters on high levels. In this case, the *extremal index function* $\theta : (0, \infty) \to [0, 1]$ is defined as a measure for clusters, where $\theta(h)$ is the extremal index of the sequence $(M_k(h))_{k\in\mathbb{N}}$ for every $h > 0$. The function $\theta(\cdot)$ is increasing. In our context we say that a continuous-time process has *extremal clusters*, if $\lim_{h\downarrow 0} \theta(h) < 1$, and otherwise it has no clusters, i. e. $\theta(h) = 1$ for every $h > 0$, by the monotony of θ. The interpretation of an extremal cluster in continuous-time is the same as in discrete-time, i.e., a continuous-time process clusters if given a large observation at some time t, there is a positive probability that any of the neighboring observations is also large.

2.3 Extensions

At the end we also want to describe the way in which it is in mathematical terms possible to investigate the locations and heights of local maxima. One possibility is by marked point processes (cf. Daley and Vere-Jones (2003), Kallenberg (1997) and Resnick (1987)). In our case, a marked point process

counts the number of elements in the set

$$\{k : a_n^{-1}(M_k(h) - b_n) \in B_0, a_n^{-1}(X_{k+t_1} - b_n) \in B_1, \ldots, a_n^{-1}(X_{k+t_l} - b_n) \in B_l\} \tag{6}$$

for any Borel sets B_j in $\mathrm{supp}(G)$, $j = 0, \ldots, l$, fixed $l \in \mathbb{N}$ and $k + t_1, \ldots, k + t_l \geq 0$, $n \in \mathbb{N}$. But there are slightly different ways to define them; see also Leadbetter et al. (1983) and Rootzén (1978). In this way, we find the locations of high level exceedances if $M_k(h)$ is large, and we describe the behavior of the process if it is on a high level by taking the limit as $n \to \infty$ in (6). More on this idea of marked point processes for Gaussian processes can be found under the name *Slepian model* going back to Lindgren in a series of papers (cf. the survey Lindgren (1984)), where $a_n^{-1}(M_k(h) - b_n)$ is replaced by an upcrossing; i. e. an *upcrossing* of level u is a point t_0 for which $X_t < u$ when $t \in (t_0 - \epsilon, t_0)$ and $X_t \geq u$ when $t \in (t_0, t_0 + \epsilon)$ for some $\epsilon > 0$. These ideas have been extended to non-Gaussian models. We refer to the very readable review paper of Leadbetter and Spaniolo (2004) on this topic and on the intensity of upcrossings on high levels. However, upcrossings have the disadvantage that there may be infinitely many in a finite time interval, so that the marked point processes converge to a degenerate limit as $n \to \infty$.

3 The Generalized Ornstein-Uhlenbeck (GOU)–Model

Generalized Ornstein-Uhlenbeck processes are applied in various areas as, e. g., in financial and insurance mathematics or mathematical physics; we refer to Carmona et al. (1997, 2001) and Donati-Martin et al. (2001) for an overview of applications. In the financial context, generalized Ornstein-Uhlenbeck processes are used as stochastic volatility models (cf. Barndorff-Nielsen and Shephard (2001a, 2001b), Barndorff-Nielsen et al. (2002)) and as insurance risk models (cf. Paulsen (1993), Klüppelberg and Kostadinova (2008), Kostadinova (2007)).

We assume throughout that $(X_t)_{t \geq 0}$ is a measurable, stationary càdlàg (right-continuous with left limits) version of the GOU-process as in (1) and that $\mathbb{P}(\sup_{0 \leq t \leq 1} |X_t| < \infty) = 1$. For two functions, f and g, we write $f(x) \sim g(x)$ as $x \to \infty$, if $\lim_{x \to \infty} f(x)/g(x) = 1$. Two distribution functions, F and H, are called *tail-equivalent* if both have support unbounded to the right and there exists some $c > 0$ such that $\lim_{x \to \infty} \overline{F}(x)/\overline{H}(x) = c$.

3.1 The Ornstein–Uhlenbeck process

Let $(X_t)_{t\geq 0}$ be a stationary GOU-process as in (1) with $\xi_t = \lambda t$ for some $\lambda > 0$; then the GOU-process reduces to a classical Ornstein-Uhlenbeck (OU) process

$$X_t = \mathrm{e}^{-\lambda t} \int_0^t \mathrm{e}^{\lambda s}\, d\eta_s + \mathrm{e}^{-\lambda t} X_0, \quad t \geq 0. \tag{7}$$

A stationary version of (7) exists if and only if $\int_{\{|x|>1\}} \log(1+|x|)\, \Pi_\eta(dx) < \infty$, where Π_η is the Lévy measure of $(\eta_t)_{t\geq 0}$. This result goes back to Wolfe (1982); see also the monograph of Sato (1999). The OU-process is a popular volatility model as introduced by Barndorff-Nielsen and Shephard (2001b); see also Shephard and Andersen (2008) in this volume.

In this paper, we study only distribution functions F of η_1 belonging to the *class of convolution equivalent distributions* denoted by $\mathcal{S}(\gamma)$ for some $\gamma \geq 0$, i. e., functions which satisfy

(i) $F(x) < 1$ for every $x \in \mathbb{R}$.
(ii) $\lim_{x\to\infty} \overline{F}(x+y)/\overline{F}(x) = \exp(-\gamma y)$ for all $y \in \mathbb{R}$ locally uniformly.
(iii) $\lim_{x\to\infty} \overline{F*F}(x)/\overline{F}(x)$ exists and is finite.

The class $\mathcal{S}(0)$ is called the class of *subexponential distributions*. For details and further references see Embrechts et al. (1997) and Watanabe (2008). An important family in $\mathcal{S}(\gamma)$ are distribution functions with tail

$$\overline{F}(x) \sim x^{-\beta} \mathrm{e}^{-\gamma x - c x^p}, \quad x \to \infty,$$

where γ, $c \geq 0$, $p < 1$, and if $c = 0$, $\beta > 1$ (cf. Klüppelberg (1989), Theorem 2.1, or Pakes (2004), Lemma 2.3). There are certain subclasses of generalized inverse Gaussian distributions, normal inverse Gaussian distributions, generalized hyperbolic distributions and CGMY distributions in $\mathcal{S}(\gamma)$, which are used for modelling financial time series (cf. Schoutens (2003)).

We investigate two different kinds of OU-models.

(M1) OU-model with $\eta_1 \in \mathcal{S}(\gamma) \cap \mathrm{MDA}(\Lambda)$. *Let $(X_t)_{t\geq 0}$ be a stationary OU-process as in (7). We assume that the distribution of η_1 is in $\mathcal{S}(\gamma) \cap \mathrm{MDA}(\Lambda)$ for some $\gamma \geq 0$.*

This assumption is sufficient for the existence of a stationary version of an OU-process. The following Proposition (cf. Proposition 2 and Proposition 3 of Fasen et al. (2006)) describes the tail behavior of X_t. The proof of this result is based on the asymptotic equivalence of the tail of the distribution function and the tail of its Lévy measure for every infinitely divisible convolution equivalent distribution in $\mathcal{S}(\gamma)$, and the representation of the Lévy measure of X_t (cf. Wolfe (1982), Theorem 2 and the monograph of Sato (1999)) as

$$\nu_X(dx) = \frac{\nu(x,\infty)}{x}\, dx \qquad \text{for } x > 0.$$

Proposition 1 *Let $(X_t)_{t\geq 0}$ be as in (M1). Then*

$$\mathbb{P}(X_t > x) = o(\mathbb{P}(\eta_1 > x)) \qquad \text{as } x \to \infty$$

and $X_t \in \mathcal{S}(\gamma) \cap \mathrm{MDA}(\Lambda)$.

This result shows that the driving Lévy process and the OU-process are in the same maximum domain of attraction, but they are not tail-equivalent. The precise relationship is given in Fasen et al. (2006). In the next model this will be different.

(M2) OU-model with $\boldsymbol{\eta_1 \in \mathcal{R}_{-\alpha}}$. *Let $(X_t)_{t\geq 0}$ be a stationary OU-process as in (7). We assume that η_1 has a regularly varying right tail distribution function, written as $\eta_1 \in \mathcal{R}_{-\alpha}$, i. e.,*

$$\mathbb{P}(\eta_1 > x) = l(x)x^{-\alpha}, \quad x \geq 0, \tag{8}$$

where $l(\cdot)$ is a slowly varying function; for more details on regular variation see Section 4 of Davis and Miksoch (2008) in this volume.

Under these assumptions there exists again a stationary version of the OU-process. All distribution functions with regularly varying tails are in $\mathcal{S}(0)$ and belong to $\mathrm{MDA}(\Phi_\alpha)$, $\alpha > 0$. In particular this means that the distribution of η_1 is also in $\mathrm{MDA}(\Phi_\alpha)$. The same techniques to compute the tail behavior of X_t as in (M1), where the tail of the Lévy measure and the probability measure are compared, are also used to derive the tail behavior of X_t in (M2) (cf. Fasen et al. (2006), Proposition 3.2).

Proposition 2 *Let $(X_t)_{t\geq 0}$ be as in (M2). Then*

$$\mathbb{P}(X_t > x) \sim (\alpha\lambda)^{-1}\mathbb{P}(\eta_1 > x) \qquad \text{as } x \to \infty.$$

This result shows that the tail of X_t is again regularly varying of index $-\alpha$, and hence, also X_t is in $\mathrm{MDA}(\Phi_\alpha)$.

3.2 The non–Ornstein–Uhlenbeck process

The last model we investigate in this paper is again a GOU-model as in (1), but it excludes the classical OU process as in (7).

(M3) Non-OU model. *Let $(X_t)_{t\geq 0}$ be a stationary GOU-model as in (1). Let $(\eta_t)_{t\geq 0}$ be a subordinator, i. e., a Lévy process with nondecreasing sample paths, and if $(\xi_t)_{t\geq 0}$ is of finite variation, then we assume additionally that*

either the drift of $(\xi_t)_{t\geq 0}$ *is non-zero, or that there is no* $r > 0$ *such that the Lévy measure of* $(\xi_t)_{t\geq 0}$ *is concentrated on* $r\boldsymbol{Z}$*. Furthermore, we suppose*

$$\mathbb{E}(\mathrm{e}^{-\alpha\xi_1}) = 1 \quad \text{for some } \alpha > 0. \tag{9}$$

We assume, finally, the moment conditions

$$\mathbb{E}|\eta_1|^{q\max\{1,d\}} < \infty \quad \text{and} \quad \mathbb{E}(\mathrm{e}^{-\max\{1,d\}p\xi_1}) < \infty \tag{10}$$

for some $d > \alpha$ *and* $p, q > 0$ *with* $1/p + 1/q = 1$.

In the classical OU-model condition (9) is not satisfied. As in many studies like the GARCH-model, Lindner and Maller (2005) apply the results of Kesten (1973) and Goldie (1991) for stochastic recurrence equations to deduce the stationarity and the heavy-tailed behavior of model (M3). In this context the stochastic recurrence equation has the form

$$X_{t+1} = A_{t+1}X_t + B_{t+1}\,, \quad t \geq 0,$$

where

$$A_t = \mathrm{e}^{-(\xi_t - \xi_{t-1})} \quad \text{and} \quad B_t = \mathrm{e}^{-\xi_t}\int_{t-1}^{t} \mathrm{e}^{\xi_{s-}}\, d\eta_s\,, \quad t \geq 0.$$

The result for the tail behavior as presented in Lindner and Maller (2005), Theorem 4.5, is the following.

Proposition 3 *Let* $(X_t)_{t\geq 0}$ *be as in (M3). Then for some* $C > 0$,

$$\mathbb{P}(X_t > x) \sim Cx^{-\alpha} \quad \text{as } x \to \infty.$$

Typical examples which satisfy (M3) are the volatility process of a continuous-time GARCH$(1,1)$ (COGARCH$(1,1)$) model introduced by Klüppelberg et al. (2004, 2006) and the volatility process of Nelson's diffusion limit of a GARCH$(1,1)$-model Nelson (1990).

Example 1 (COGARCH$(1,1)$ process) The right-continuous version of the volatility process of the COGARCH$(1,1)$ process is defined as GOU-process as in (1), where

$$\xi_t = ct - \sum_{0<s\leq t} \log(1 + \beta\mathrm{e}^{c}(\Delta L_s)^2) \quad \text{and} \quad \eta_t = \lambda t \quad \text{for } t \geq 0,$$

$\lambda, c > 0$, $\beta \geq 0$ are constants and $(L_t)_{t\geq 0}$ is a Lévy process (cf. Lindner (2008) of this volume). The assumptions in (M3) are satisfied if and only if

$$-\alpha c + \int ((1 + \beta\mathrm{e}^{c}y^2)^{\alpha} - 1)\, \Pi_L(dy) \quad \text{and} \quad \mathbb{E}|L_1|^{2\widetilde{d}} < \infty \text{ for some } \widetilde{d} > \alpha,$$

where Π_L denotes the Lévy measure of L.

Example 2 (Nelson's diffusion model) The Nelson's diffusion model, originally defined as solution of the stochastic differential equation

$$dX_t = \lambda(a - X_t)\, dt + \sigma X_t\, dB_t,$$

where $a, \lambda, \sigma > 0$ and $(B_t)_{t\geq 0}$ is a Brownian motion, is by Theorem 52 on p. 328 in Protter (2004) a GOU-process with

$$\xi_t = -\sigma B_t + \left(\frac{1}{2}\sigma^2 + \lambda\right) t \quad \text{and} \quad \eta_t = \lambda a t \quad \text{for } t \geq 0.$$

Since

$$\mathbb{E}(\mathrm{e}^{-u\xi_1}) = \exp\left(\frac{1}{2}\sigma^2 u^2 - \left(\frac{1}{2}\sigma^2 + \lambda\right) u\right)$$

we have $\mathbb{E}(\mathrm{e}^{-\alpha\xi_1}) = 1$ for $\alpha = 1 + 2\lambda/\sigma^2$.

For more details on these examples we refer to Lindner (2008) in this volume.

3.3 Comparison of the models

At first glance, the results presented in Propositions 1-3 are surprising. We start with a comparison of models (M1) and (M3) driven by the same Lévy process $(\eta_t)_{t\geq 0}$. In model (M1), the tail of η_1 is heavier than the tail of X_t. In contrast, in model (M3) the existence of the $q\alpha$ moment of η_1 by (10) results in the tail of η_1 being at most $-q\alpha$ regularly varying and hence, lighter tailed than X_t. Taking now, in models (M1) and (M3), the same Lévy process $(\eta_t)_{t\geq 0}$, it ensures that X_t has a different tail behavior in each model. In (M1) it is lighter tailed and in (M3) it is heavier tailed than η_1.

Next we compare the OU-models (M1) and (M2), which have the same Lévy process $(\xi_t)_{t\geq 0}$. In model (M1) we find that X_t is lighter tailed than η_1; in (M2) we find the tail-equivalence of the distribution function of X_t and η_1.

We conclude that both Lévy processes $(\xi_t)_{t\geq 0}$ and $(\eta_t)_{t\geq 0}$ are contributing factors to the tail behavior of X_t.

4 Tail Behavior of the Sample Maximum

It is the tail of the distribution of the sample maximum $M(h)$ as in (2) for some $h > 0$, rather than X_t itself, that determines the limit distribution

of the normalized process $M(T)$ as $T \to \infty$. The tail behavior of $M(h)$ is affected differently in models (M1)–(M3).

For (M1) the derivation of the asymptotic behavior of the tail of $M(h)$ is much more involved than for (M2) and given in Fasen (2008b). For model (M2) the following asymptotic behavior holds:

$$\begin{aligned}
&\mathbb{P}(M(h) > x) \\
&= \mathbb{P}\left(\sup_{0\le t\le h}\left\{\mathrm{e}^{-\lambda t}\int_0^t \mathrm{e}^{\lambda s}\,d\eta_s + \mathrm{e}^{-\lambda t}X_0\right\} > x\right) \\
&\sim \mathbb{P}\left(\int_0^h \sup_{0\le t\le h}\left\{\mathbf{1}_{[0,t)}\,\mathrm{e}^{-\lambda(t-s)}\right\}\,d\eta_s > x\right) + \mathbb{P}\left(\sup_{0\le t\le h}\left\{\mathrm{e}^{-\lambda t}\right\}X_0 > x\right) \\
&= \mathbb{P}(\eta_h > x) + \mathbb{P}(X_0 > x) \quad \text{as } x \to \infty.
\end{aligned}$$

The mathematical proof (see Fasen (2005), Proposition 3.2) is based on results of Rosiński and Samorodnitsky (1993) investigating the tail behavior of random variables in $\mathcal{S}(0)$, which are functionals acting on infinitely divisible processes. In model (M3) only the last summand of representation (7) influences the tail behavior, since

$$\mathbb{E}\left|\sup_{0\le t\le h}\mathrm{e}^{-\xi_t}\int_0^t \mathrm{e}^{\xi_{s-}}\,d\eta_s\right|^d < \infty$$

(cf. Fasen (2008a), Remark 2.3 *(iii)*). Hence, Klüppelberg et al. (2006), Lemma 2, and Breiman (1965), Proposition 3, give as $x \to \infty$,

$$\begin{aligned}
\mathbb{P}(M(h) > x) &= \mathbb{P}\left(\sup_{0\le t\le h}\left\{\mathrm{e}^{-\xi_t}\int_0^t \mathrm{e}^{\xi_{s-}}\,d\eta_s + \mathrm{e}^{-\xi_t}X_0\right\} > x\right) \\
&\sim \mathbb{P}\left(\sup_{0\le t\le h}\left\{\mathrm{e}^{-\xi_t}\right\}X_0 > x\right) \\
&\sim \mathbb{E}\left(\sup_{0\le s\le h}\mathrm{e}^{-\alpha\xi_s}\right)\mathbb{P}(X_0 > x).
\end{aligned}$$

We summarize the tail behavior of $M(h)$ for the different models.

Proposition 4

(a) OU-model with $\eta_1 \in \mathcal{S}(\gamma) \cap \mathrm{MDA}(\Lambda)$ as in (M1):

$$\mathbb{P}(M(h) > x) \sim h\frac{\mathbb{E}(\mathrm{e}^{\gamma X_0})}{\mathbb{E}(\mathrm{e}^{\gamma\eta_1})}\mathbb{P}(\eta_1 > x) \quad \text{as } x \to \infty.$$

(b) OU-model with $\eta_1 \in \mathcal{R}_{-\alpha}$ as in (M2):

$$\mathbb{P}(M(h) > x) \sim \left(h + \frac{1}{\alpha\lambda}\right)\mathbb{P}(\eta_1 > x) \quad \text{as } x \to \infty.$$

(c) Non-OU model as in (M3):

$$\mathbb{P}(M(h) > x) \sim \mathbb{E}\left(\sup_{0\le s\le h} \mathrm{e}^{-\alpha\xi_s}\right)\mathbb{P}(X_t > x) \quad \text{as } x \to \infty.$$

In all three models $M(h)$ is in the same maximum domain of attraction as X_t.

5 Running sample Maxima and Extremal Index Function

The classic problem arising from studying the extremal behavior of stochastic processes is the asymptotic behavior of the sample maxima $M(T)$ as $T \to \infty$. One of the first researchers turning from the extremal behavior of Gaussian processes to stable processes was Rootzén (1978). His results already include the asymptotic behavior of sample maxima of OU-processes driven by stable Lévy motions and their marked point process behavior, where the definition of the marked point process is slightly different to Section 2.3. Generalizations of his results to regularly varying processes including model (M2) are presented in Fasen (2005). Model (M1) was investigated in Fasen et al. (2006), but more details can be found in Fasen (2008b). A proof of the asymptotic behavior in model (M3) is given in Fasen (2008a). We denote by $x^+ = \max\{0, x\}$ for $x \in \mathbb{R}$.

Proposition 5

(a) OU-model with $\eta_1 \in \mathcal{S}(\gamma) \cap \mathrm{MDA}(\Lambda)$ as in (M1):
Let $a_T > 0, b_T \in \mathbb{R}$ be sequences of constants such that

$$\lim_{T\to\infty} T\mathbb{P}(M(1) > a_T x + b_T) = \exp(-x) \quad \text{for } x \in \mathbb{R}.$$

Then

$$\lim_{T\to\infty} \mathbb{P}(a_T^{-1}(M(T) - b_T) \le x) = \exp(-\mathrm{e}^{-x}) \quad \text{for } x \in \mathbb{R},$$

and

$$\theta(h) = 1 \quad \text{for } h > 0.$$

(b) OU-model with $\eta_1 \in \mathcal{R}_{-\alpha}$ as in (M2):
Let $a_T > 0$ be a sequence of constants such that

$$\lim_{T\to\infty} T\mathbb{P}(M(1) > a_T x) = x^{-\alpha} \quad \text{for } x > 0.$$

Then

$$\lim_{T\to\infty} \mathbb{P}(a_T^{-1} M(T) \le x) = \exp\left(-\frac{\alpha\lambda}{\alpha\lambda+1} x^{-\alpha}\right) \quad \textit{for } x > 0,$$

and

$$\theta(h) = \frac{h\alpha\lambda}{h\alpha\lambda+1} \quad \textit{for } h > 0.$$

(c) Non-OU model as in (M3):
Let $a_T > 0$ be a sequence of constants such that

$$\lim_{T\to\infty} T\mathbb{P}(M(1) > a_T x) = x^{-\alpha} \quad \textit{for } x > 0.$$

Then

$$\lim_{T\to\infty} \mathbb{P}(a_T^{-1} M(T) \le x) = \exp\left(-\frac{\mathbb{E}\big(\sup_{0\le s\le 1} \mathrm{e}^{-\alpha\xi_s} - \sup_{s\ge 1} \mathrm{e}^{-\alpha\xi_s}\big)^+}{\mathbb{E}\left(\sup_{0\le s\le 1} \mathrm{e}^{-\alpha\xi_s}\right)} x^{-\alpha}\right)$$

for $x > 0$, and

$$\theta(h) = h\frac{\mathbb{E}\left(\sup_{0\le s\le 1} \mathrm{e}^{-\alpha\xi_s} - \sup_{s\ge 1} \mathrm{e}^{-\alpha\xi_s}\right)^+}{\mathbb{E}\left(\sup_{0\le s\le h} \mathrm{e}^{-\alpha\xi_s}\right)} \quad \textit{for } h > 0.$$

These results reflect the fact that model (M1) has no clusters of extremes on high levels, whereas both regularly varying models (M2) and (M3) have them. In particular, models (M2) and (M3) do not satisfy the anti-cluster condition C'.

For the behavior of the marked point processes of model (M1) and (M2) we refer to Fasen et al. (2006) and of (M3) to Fasen (2008a).

6 Conclusion

All continuous-time models $(X_t)_{t\ge 0}$ presented in Section 3 are heavy-tailed models, which model stationary continuous-time processes with jumps. The OU-model in (M1) has no clusters of extremes. This property has been confirmed so far in all investigated OU-models in MDA(Λ), including Gaussian OU-processes (cf. Albin (2008)). However, the regularly varying models (M2) and (M3) have extremal clusters on high levels.

One generalization of the OU-process is the supOU process introduced by Barndorff-Nielsen (2001), where the driving Lévy process is replaced by an infinitely divisible random measure. Modelling long range dependence, in the sense that the autocovariance function decreases very slowly, is a special feature of this class of processes. All models presented in this paper have exponentially decreasing covariance functions and do not allow long range

dependence. SupOU processes have an extremal behavior similar to that of OU-models, see Fasen and Klüppelberg (2007). This means that only regularly varying supOU processes have extremal clusters.

Another extension of OU-processes are continuous-time ARMA (CARMA) processes, as presented in Brockwell (2008) of this volume. In such models the exponentially decreasing kernel function of an OU-process is replaced by a more general kernel function. The results of Fasen (2008b, 2005) show that a CARMA process and an OU-process, driven by the same Lévy process, have similar extremal behavior. The regularly varying CARMA processes show extremal clusters. In the case in which the driving Lévy process of the CARMA process has marginals in $\mathcal{S}(\gamma) \cap \mathrm{MDA}(\Lambda)$, and the kernel functions have only one maximum, there are again no extremal clusters. If they have more than one maximum, then they may also model extremal clusters.

References

Albin, J. M. P. (2008): On extremes of infinitely divisible Ornstein-Uhlenbeck processes. *Preprint, available at* `http://www.math.chalmers.se/~palbin/`

Albin, J. M. P. (1987): *On Extremal Theory for Nondifferentiable Stationary Processes*. PhD thesis, Lund University.

Albin, J. M. P. (1990): On extremal theory for stationary processes. *Ann. Probab.* **18**, 92–128.

Barndorff-Nielsen, O. E. (2001): Superposition of Ornstein–Uhlenbeck type processes. *Theory Probab. Appl.* **45**, 175–194.

Barndorff-Nielsen, O. E., Nicolata, E. and Shephard, N. (2002): Some recent developments in stochastic volatility modelling. *Quantitative Finance* **2**, 11–23.

Barndorff-Nielsen, O. E. and Shephard, N. (2001a): Modelling by Lévy processes for financial econometrics. *In: Barndorff-Nielsen, O. E., Mikosch, T., and Resnick, S. I. (Eds.): Lévy Processes: Theory and Applications*, 283–318. Birkhäuser, Boston.

Barndorff-Nielsen, O. E. and Shephard, N. (2001b): Non-Gaussian OU based models and some of their uses in financial economics. *J. Roy. Statist. Soc. Ser. B* **63**, 167–241.

Berman, S. M. (1992): *Sojourns and Extremes of Stochastic Processes*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove.

Breiman, L. (1965): On some limit theorems similar to the arc-sine law. *Theory Probab. Appl.* **10**, 323–331.

Brockwell, P. (2008): Lévy driven continuous-time ARMA processes. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 456–480. Springer, New York.

Carmona, P., Petit, F. and Yor, M. (1997): On the distribution and asymptotic results for exponential Lévy processes. *In: Yor, M. (Ed.): Exponential Functionals and Principal Values Related to Brownian Motion*, 73–130. Biblioteca de le Revista Matemàtica Iberoamericana.

Carmona, P., Petit, F. and Yor, M. (2001): Exponential functionals of Lévy processes. *In: Barndorff-Nielsen, O. E., Mikosch, T., and Resnick, S. (Eds.): Lévy Processes, Theory and Applications*, 41–55. Birkhäuser, Boston.

Daley, D. J. and Vere-Jones, D. (2003): *An Introduction to the Theory of Point Processes. Vol I: Elementary Theory and Methods*, 2nd edition. Springer, New York.

Davis, R. A. and Miksoch, T. (2008): Probabilistic properties of stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 255–267. Springer, New York.

De Haan, L. and Ferreira, A. (2006): *Extreme Value Theory: An Introduction*. Springer.
Donati-Martin, C., Ghomrasni, R. and Yor, M. (2001): On certain Markov processes attached to exponential functionals of Brownian motion; application to Asian options. *Rev. Mat. Iberoamericana* **17**, 179–193.
Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997): *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.
Fasen, V. (2005): Extremes of regularly varying mixed moving average processes. *Adv. in Appl. Probab.* **37**, 993–1014.
Fasen, V. (2008a): Asymptotic results for sample ACF and extremes of integrated generalized Ornstein-Uhlenbeck processes. *Preprint, available at* `http://www.ma.tum.de/stat/`
Fasen, V. (2008b): Extremes of mixed MA processes in the class of convolution equivalent distributions. *Preprint, available at* `http://www.ma.tum.de/stat/`
Fasen, V. and Klüppelberg, C. (2007): Extremes of supOU processes. *In: Benth, F. E., Di Nunno, G., Lindstrom, T., Oksendal, B. and Zhang, T. (Eds.): Stochastic Analysis and Applications: The Abel Symposium 2005*, 340–359. Springer.
Fasen, V., Klüppelberg, C. and Lindner, A. (2006): Extremal behavior of stochastic volatility models. *In: Shiryaev, A., Grossinho, M. D. R., Oliviera, P. and Esquivel, M. (Eds.): Stochastic Finance*, 107–155. Springer, New York.
Goldie, C. M. (1991): Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.* **1**, 126–166.
Kac, M. (1943): On the average number of real roots of a random algebraic equation. *Bull. Amer. Math. Soc.* **49**, 314–320.
Kac, M. and Slepian, D. (1959): Large excursions of Gaussian processes. *Ann. Math. Statist.* **30**, 1215–1228.
Kallenberg, O. (1997): *Foundations of Modern Probability*. Springer, New York.
Kesten, H. (1973): Random difference equations and renewal theory for products of random matrices. *Acta Math.* **131**, 207–248.
Klüppelberg, C. (1989): Subexponential distributions and characterizations of related classes. *Probab. Theory Relat. Fields* **82**, 259–269.
Klüppelberg, C., Lindner, A. and Maller, R. (2004): A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour. *J. Appl. Probab.* **41**, 601–622.
Klüppelberg, C., Lindner, A. and Maller, R. (2006): Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models. *In: Kabanov, Y., Lipster, R., and Stoyanov, J. (Eds.): From Stochastic Calculus to Mathematical Finance. The Shiryaev Festschrift*, 393–419. Springer, Berlin.
Klüppelberg, K. and Kostadinova, T. (2008): Integrated insurance risk models with exponential Lévy investment. *Insurance Math. Econ.* to appear.
Kostadinova, T. (2007): Optimal investment for insurers, when stock prices follows an exponential Lévy process. *Insurance Math. Econ.* **41**, 250–263.
Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.
Leadbetter, M. R. and Rootzén, H. (1982): Extreme value theory for continuous parameter stationary processes. *Z. Wahrsch. verw. Gebiete* **60**, 1–20.
Leadbetter, M. R. and Rootzén, H. (1988): Extremal theory for stochastic processes. *Ann. Probab.* **16**, 431–478.
Leadbetter, M. R. and Spaniolo, G. V. (2004): Reflections on Rice's formulae for level crossings–history, extensions and use. *Aust. N. J. Stat.* **46**, 173–180.
Lindgren, G. (1984): Use and structure of Slepian model processes for prediction and detection in crossing and extreme value theory. *In: Oliveira, J. T. D. (Ed.): Statistical Extremes and Applications*, 261–284. Dordrecht. Reidel.
Lindner, A. (2008): Continuous time approximations to GARCH and stochastic volatility models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 481–496. Springer, New York.

Lindner, A. and Maller, R. (2005): Lévy integrals and the stationarity of generalised Ornstein-Uhlenbeck processes. *Stoch. Proc. Appl.* **115**, 1701–1722.

Maller, R., Müller, G., and Szimayer, A. (2008): Ornstein-Uhlenbeck processes and extensions. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 420–437. Springer, New York.

Nelson, D. B. (1990): ARCH models as diffusion approximations. *J. Econometrics* **45**, 7–38.

Pakes, A. G. (2004): Convolution equivalence and infinite divisibility. *J. Appl. Probab.* **41**, 407–424.

Paulsen, J. (1993): Risk theory in a stochastic economic environment. *Stoch. Proc. Appl.* **46**, 327–361.

Protter, P. E. (2004): *Stochastic Integration and Differential Equations*, 2nd edition. Springer, Berlin.

Resnick, S. I. (1987): *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.

Rice, S. O. (1939): The distribution of the maximum of a random curve. *Amer. J. Math.* **61**, 409–416.

Rice, S. O. (1944): Mathematical analysis of random noise. *Bell. System Techn. J.* **23**, 282–332.

Rice, S. O. (1945): Mathematical analysis of random noise. *Bell. System Techn. J.* **24**, 46–156.

Rootzén, H. (1978): Extremes of moving averages of stable processes. *Ann. Probab.* **6**, 847–869.

Rosiński, J. and Samorodnitsky, G. (1993): Distributions of subadditive functionals of sample paths of infinitely divisible processes. *Ann. Probab.* **21**, 996–1014.

Sato, K. (1999): *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.

Schoutens, W. (2003): *Lévy Processes in Finance*. Wiley, Chichester.

Shephard, N. and Andersen, T.G. (2008): Stochastic volatility: origins and overview. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 233–254. Springer, New York.

Slepian, D. (1961): First passage time for particular Gaussian processes. *Ann. Math . Stat.* **32**, 610–612.

Slepian, D. (1962): The one-sided barrier problem for Gaussian noise. *Bell System Techn. J.* **41**, 463–501.

Volkonskii, V. A. and Rozanov, Y. A. (1959): Some limit theorems for random functions, I. *Theory Probab. Appl.* **4**, 178–197.

Volkonskii, V. A. and Rozanov, Y. A. (1961): Some limit theorems for random functions, II. *Theory Probab. Appl.* **6**, 186–198.

Watanabe, T. (2008): Convolution equivalence and distributions of random sums. *Probab. Theory Relat. Fields* to appear.

Wolfe, S. J. (1982): On a continuous analogue of the stochastic difference equation $x_n = \rho x_{n-1} + b_n$. *Stoch. Proc. Appl.* **12**, 301–312.

Cointegration: Overview and Development

Søren Johansen

Abstract This article presents a survey of the analysis of cointegration using the vector autoregressive model. After a few illustrative economic examples, the three model based approaches to the analysis of cointegration are discussed. The vector autoregressive model is defined and the moving average representation of the solution, the Granger representation, is given. Next the interpretation of the model and its parameters and likelihood based inference follows using reduced rank regression. The asymptotic analysis includes the distribution of the Gaussian maximum likelihood estimators, the rank test, and test for hypotheses on the cointegrating vectors. Finally, some applications and extensions of the basic model are mentioned and the survey concludes with some open problems.

1 Introduction

Granger (1983) coined the term cointegration as a formulation of the phenomenon that nonstationary processes can have linear combinations that are stationary. It was his investigations of the relation between cointegration and error correction that brought modeling of vector autoregressions with unit roots and cointegration to the center of attention in applied and theoretical econometrics; see Engle and Granger (1987).

During the last 20 years, many have contributed to the development of theory and applications of cointegration. The account given here focuses on theory, more precisely on likelihood based theory for the vector autoregressive model and its extensions; see Johansen (1996). By building a statistical model

Søren Johansen
Department of Applied Mathematics and Statistics, University of Copenhagen, e-mail: sjo@math.ku.dk

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_29,

as a framework for inference, one has to make explicit assumptions about the model used and hence has a possibility of checking the assumptions made.

1.1 Two examples of cointegration

As a first simple economic example of the main idea in cointegration, consider the exchange rate series, e_t, between Australian and US dollars and four time series p_t^{au}, p_t^{us}, i_t^{au}, i_t^{us} : log consumer price and five year treasury bond rates in Australia and US. If the quarterly series from 1972:1 to 1991:1 are plotted, they clearly show nonstationary behavior, and we discuss in the following a method of modeling such nonstationary time series. As a simple example of an economic hypothesis consider Purchasing Power Parity (PPP), which asserts that $e_t = p_t^{us} - p_t^{au}$. This identity is not found in the data, so a more realistic formulation is that $ppp_t = e_t - p_t^{us} + p_t^{au}$ is a stationary process, possibly with mean zero. Thus we formulate the economic relation, as expressed by PPP, as a stationary relation among nonstationary processes. The purpose of modeling could be to test the null hypothesis that ppp_t is stationary, or in other words that $(e_t, p_t^{us}, p_t^{au} i_t^{au}, i_t^{us})$ cointegrate with $(1, -1, 1, 0, 0)'$ as a cointegration vector. If that is not found, an outcome could be to suggest other cointegration relations, which have a better chance of capturing co-movements of the five processes in the information set. For a discussion of the finding that real exchange rate, ppp_t, and the spread, $i_t^{au} - i_t^{us}$, are cointegrated $I(1)$ processes so that a linear combination $ppp_t - c\,(i_t^{au} - i_t^{us})$ is stationary, see Juselius and MacDonald (2004).

Another example is one of the first applications of the idea of cointegration in finance; see Campell and Shiller (1987). They considered a present value model for the price of a stock Y_t at the end of period t and the dividend y_t paid during period t. They assume that there is a vector autoregressive model describing the data which contain Y_t and y_t and may contain values of other financial assets. The expectations hypothesis is expressed as

$$Y_t = \theta(1-\delta)\sum_{i=0}^{\infty} \delta^j E_t y_{t+i} + c,$$

where c and θ are positive constants and the discount factor δ is between 0 and 1. The notation $E_t y_{t+i}$ means model based conditional expectations of y_{t+i} given information in the data at the end of period t. By subtracting θy_t, the model is written as

$$Y_t - \theta y_t = \theta(1-\delta)\sum_{i=0}^{\infty} \delta^j E_t(y_{t+i} - y_t) + c.$$

It is seen that when the processes y_t and Y_t are nonstationary and their differences stationary, the present value model implies that the right hand side and hence the left hand side are stationary. Thus there is cointegration between Y_t and y_t with a cointegration vector $\beta' = (1, -\theta, 0, \ldots, 0)$; see section 6.1 for a discussion of rational expectations and cointegration.

There are at present three different ways of modeling the cointegration idea in a parametric statistical framework. To illustrate the ideas they are formulated in the simplest possible case, leaving out deterministic terms.

1.2 Three ways of modeling cointegration

1.2.1 The regression formulation

The multivariate process $x_t = (x'_{1t}, x'_{2t})'$ of dimension $p = p_1 + p_2$ is given by the regression equations

$$\begin{aligned} x_{1t} &= \gamma' x_{2t} + u_{1t}, \\ \Delta x_{2t} &= u_{2t}, \end{aligned}$$

where $u_t = (u'_{1t}, u'_{2t})'$ is a linear invertible process defined by i.i.d. errors ε_t with mean zero and finite variance. The assumptions behind this model imply that x_{2t} is nonstationary and not cointegrated, and hence the cointegration rank, p_1, is known so that models for different ranks are not nested. The first estimation method used in this model is least squares regression, Engle and Granger (1987), which is shown to give a superconsistent estimator by Stock (1987). This estimation method gives rise to residual based tests for cointegration. It was shown by Phillips and Hansen (1990) that a modification of the regression estimator, involving a correction using the long-run variance of the process u_t, would give useful methods for inference for coefficients of cointegration relations; see also Phillips (1991).

1.2.2 The autoregressive formulation

The autoregressive formulation is given by

$$\Delta x_t = \alpha \beta' x_{t-1} + \varepsilon_t,$$

where ε_t are i.i.d. errors with mean zero and finite variance, and α and β are $p \times r$ matrices of rank r. Under the condition that Δx_t is stationary, the solution is

$$x_t = C \sum_{i=1}^{t} \varepsilon_i + \sum_{i=0}^{\infty} C_i \varepsilon_{t-i} + A, \tag{1}$$

where $C = \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\alpha'_\perp$ and $\beta' A = 0$. Here $\beta_\perp$ is a full rank $p \times (p - r)$ matrix so that $\beta'\beta_\perp = 0$. This formulation allows for modeling of both the long-run relations, $\beta' x$, and the adjustment, or feedback α, towards the attractor set $\{x : \beta' x = 0\}$ defined by the long-run relations. Models for different cointegration ranks are nested and the rank can be analyzed by likelihood ratio tests. Thus the model allows for a more detailed description of the data than the regression model. Methods usually applied for the analysis are derived from the Gaussian likelihood function, which are discussed here; see also Johansen (1988, 1996), and Ahn and Reinsel (1990).

1.2.3 The unobserved component formulation

Let x_t be given by

$$x_t = \xi\eta' \sum_{i=1}^{t} \varepsilon_i + v_t,$$

where v_t is a linear process, typically independent of the process ε_t, which is i.i.d. with mean zero and finite variance.

In this formulation too, hypotheses of different ranks are nested. The parameters are linked to the autoregressive formulation by $\xi = \beta_\perp$ and $\eta = \alpha_\perp$, even though the linear process in (1) depends on the random walk part, so the unobserved components model and the autoregressive model are not the same. Thus both adjustment and cointegration can be discussed in this formulation, and hypotheses on the rank can be tested. Rather than testing for unit roots one tests for stationarity, which is sometimes a more natural formulation. Estimation is usually performed by the Kalman filter, and asymptotic theory of the rank tests has been worked out by Nyblom and Harvey (2000).

1.3 The model analyzed in this article

In this article cointegration is modelled by the vector autoregressive model for the $p-$dimensional process x_t

$$\Delta x_t = \alpha\,(\beta' x_{t-1} + \Upsilon D_t) + \sum_{i=1}^{k-1} \Gamma_i\, \Delta x_{t-i} + \Phi d_t + \varepsilon_t, \tag{2}$$

where ε_t are i.i.d. with mean zero and variance Ω, and D_t and d_t are deterministic terms, like constant, trend, seasonal- or intervention dummies. The matrices α and β are $p \times r$ where $0 \le r \le p$. The parametrization of the deterministic term $\alpha\Upsilon D_t + \Phi d_t$, is discussed in section 2.2. Under suitable conditions, see again section 2.2, the processes $\beta' x_t$ and Δx_t are stationary

around their means, and (2) can be formulated as

$$\Delta x_t - E(\Delta x_t) = \alpha\left(\beta' x_{t-1} - E(\beta' x_{t-1})\right) + \sum_{i=1}^{k-1} \Gamma_i \left(\Delta x_{t-i} - E(\Delta x_{t-i})\right) + \varepsilon_t.$$

This shows how the change of the process reacts to feedback from disequilibrium errors $\beta' x_{t-1} - E(\beta' x_{t-1})$ and $\Delta x_{t-i} - E(\Delta x_{t-i})$, via the short-run adjustment coefficients α and Γ_i. The equation $\beta' x_t - E(\beta' x_t) = 0$ defines the long-run relations between the processes.

There are many surveys of the theory of cointegration; see for instance Watson (1994) or Johansen (2006a). The topic has become part of most textbooks in econometrics; see among others Banerjee et al. (1993), Hamilton (1994), Hendry (1995) and Lütkepohl (2006). For a general account of the methodology of the cointegrated vector autoregressive model, see Juselius (2006).

2 Integration, Cointegration and Granger's Representation Theorem

The basic definitions of integration and cointegration are given together with a moving average representation of the solution of the error correction model (2). This solution reveals the stochastic properties of the solution. Finally the interpretation of cointegration relations is discussed.

2.1 Definition of integration and cointegration

The vector autoregressive model for the $p-$dimensional process x_t given by (2) is a dynamic stochastic model for all components of x_t. By recursive substitution, the equations define x_t as function of initial values, $x_0, \ldots, x_{-k+1}$, errors $\varepsilon_1, \ldots, \varepsilon_t$, deterministic terms, and parameters. Properties of the solution of these equations are studied through the characteristic polynomial

$$\Psi(z) = (1-z)I_p - \Pi z - (1-z)\sum_{i=1}^{k-1} \Gamma_i z^i \tag{3}$$

with determinant $|\Psi(z)|$. The function $C(z) = \Psi(z)^{-1}$ has poles at the roots of the polynomial $|\Psi(z)|$ and the position of the poles determines the stochastic properties of the solution of (2). First a well known result is mentioned; see Anderson (1984).

Theorem 1 *If $|\Psi(z)| = 0$ implies that $|z| > 1$, then α and β have full rank p, and the coefficients of $\Psi^{-1}(z) = \sum_{i=0}^{\infty} C_i z^i$ are exponentially decreasing. Let $\mu_t = \sum_{i=0}^{\infty} C_i(\alpha\Upsilon D_{t-i} + \Phi d_{t-i})$. Then the distribution of the initial values of x_t can be chosen so that $x_t - \mu_t$ is stationary. Moreover, x_t has the moving average representation*

$$x_t = \sum_{i=0}^{\infty} C_i \varepsilon_{t-i} + \mu_t. \tag{4}$$

Thus the exponentially decreasing coefficients are found by simply inverting the characteristic polynomial if the roots are outside the unit disk. If this condition fails, the equations generate nonstationary processes of various types, and the coefficients are not exponentially decreasing. Still, the coefficients of $C(z)$ determine the stochastic properties of the solution of (2), as is discussed in section 2.2. A process of the form (4) is a linear process and forms the basis for the definitions of integration and cointegration.

Definition 1 The process x_t is integrated of order 1, $I(1)$, if $\Delta x_t - E(\Delta x_t)$ is a linear process, with $C(1) = \sum_{i=0}^{\infty} C_i \neq 0$. If there is a vector $\beta \neq 0$ so that $\beta' x_t$ is stationary around its mean, then x_t is cointegrated with cointegration vector β. The number of linearly independent cointegration vectors is the cointegration rank.

Example 1 A bivariate process is given for $t = 1, \ldots, T$ by the equations

$$\begin{aligned} \Delta x_{1t} &= \alpha_1(x_{1t-1} - x_{2t-1}) + \varepsilon_{1t}, \\ \Delta x_{2t} &= \alpha_2(x_{1t-1} - x_{2t-1}) + \varepsilon_{2t}. \end{aligned}$$

Subtracting the equations, we find that the process $y_t = x_{1t} - x_{2t}$ is autoregressive and stationary if $|1 + \alpha_1 - \alpha_2| < 1$ and the initial value is given by its invariant distribution. Similarly we find that $S_t = \alpha_2 x_{1t} - \alpha_1 x_{2t}$ is a random walk, so that

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \frac{1}{\alpha_2 - \alpha_1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} S_t - \frac{1}{\alpha_2 - \alpha_1} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} y_t.$$

This shows, that when $|1 + \alpha_1 - \alpha_2| < 1$, x_t is $I(1)$, $x_{1t} - x_{2t}$ is stationary, and $\alpha_2 x_{1t} - \alpha_1 x_{2t}$ is a random walk, so that x_t is a cointegrated $I(1)$ process with cointegration vector $\beta' = (1, -1)$. We call S_t a common stochastic trend and α the adjustment coefficients.

Example 1 presents a special case of the Granger Representation Theorem, which gives the moving average representation of the solution of the error correction model.

2.2 The Granger Representation Theorem

If the characteristic polynomial $\Psi(z)$ defined in (3) has a unit root, then $\Psi(1) = -\Pi$ is singular, of rank $r < p$, and the process is not stationary. Next the $I(1)$ condition is formulated. It guarantees that the solution of (2) is a cointegrated $I(1)$ process. Let $\Gamma = I_p - \sum_{i=1}^{k-1} \Gamma_i$ and denote for a $p \times m$ matrix a by $a_\perp$ a $p \times (p-m)$ matrix of rank $p-m$.

Condition 1 (The $I(1)$ condition)
The $I(1)$condition is satisfied if $|\Psi(z)| = 0$ implies that $|z| > 1$ or $z = 1$ and that

$$|\alpha'_\perp \Gamma \beta_\perp| \neq 0. \tag{5}$$

Condition (5) is needed to avoid solutions that are integrated of order 2 or higher; see section 6. For a process with one lag $\Gamma = I_p$ and

$$\beta' x_t = (I_r + \beta'\alpha)\beta' x_{t-1} + \beta'\varepsilon_t.$$

In this case the $I(1)$ condition is equivalent to the condition that the absolute value of the eigenvalues of $I_r + \beta'\alpha$ are bounded by one, and in example 1 the condition is $|1 + \alpha_1 - \alpha_2| < 1$.

Theorem 2 (The Granger Representation Theorem)
Let $\Psi(z)$ be defined by (3). *If $\Psi(z)$ has unit roots and the $I(1)$ condition is satisfied, then*

$$(1-z)\Psi(z)^{-1} = C(z) = \sum_{i=0}^{\infty} C_i z^i = C + (1-z)C^*(z) \tag{6}$$

converges for $|z| \leq 1 + \delta$ for some $\delta > 0$ and

$$C = \beta_\perp (\alpha'_\perp \Gamma \beta_\perp)^{-1} \alpha'_\perp. \tag{7}$$

The solution x_t of equation (2) *has the moving average representation*

$$x_t = C \sum_{i=1}^{t} (\varepsilon_i + \Phi d_i) + \sum_{i=0}^{\infty} C_i^* (\varepsilon_{t-i} + \Phi d_{t-i} + \alpha \Upsilon D_{t-i}) + A, \tag{8}$$

where A depends on initial values, so that $\beta' A = 0$.

This result implies that Δx_t and $\beta' x_t$ are stationary, so that x_t is a cointegrated $I(1)$ process with r cointegration vectors β and $p-r$ common stochastic trends $\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i$. The interpretation of this is that among p nonstationary processes the model (2) generates r stationary or stable relations and $p-r$ stochastic trends or driving trends, which create the nonstationarity.

The result (6) rests on the observation that the singularity of $\Psi(z)$ for $z = 1$ implies that $\Psi(z)^{-1}$ has a pole at $z = 1$. Condition (5) is a condition

for this pole to be of order one. This is not proved here, see Johansen (2006b), but it is shown how this result can be applied to prove the representation result (8), which shows how coefficients of the inverse polynomial determine the properties of x_t.

We multiply $\Psi(L)x_t = \Phi d_t + \alpha\Upsilon D_t + \varepsilon_t$ by

$$(1-L)\Psi(L)^{-1} = C(L) = C + (1-L)C^*(L)$$

and find

$$\Delta x_t = (1-L)\Psi(L)^{-1}\Psi(L)x_t = (C + \Delta C^*(L))(\varepsilon_t + \alpha\Upsilon D_t + \Phi d_t).$$

Now define the stationary process $z_t = C^*(L)\varepsilon_t$ and the deterministic function $\mu_t = C^*(L)(\alpha\Upsilon D_t + \Phi d_t)$, and note that $C\alpha\Upsilon = 0$, so that

$$\Delta x_t = C(\varepsilon_t + \Phi d_t) + \Delta(z_t + \mu_t),$$

which cumulates to

$$x_t = C\sum_{i=1}^{t}(\varepsilon_i + \Phi d_i) + z_t + \mu_t + A,$$

where $A = x_0 - z_0 - \mu_0$. The distribution of x_0 is chosen so that $\beta' x_0 = \beta'(z_0 + \mu_0)$, and hence $\beta' A = 0$. Then x_t is $I(1)$ and $\beta' x_t = \beta' z_t + \beta'\mu_t$ is stationary around its mean $E(\beta' x_t) = \beta'\mu_t$. Finally, Δx_t is stationary around its mean $E(\Delta x_t) = C\Phi d_t + \Delta\mu_t$.

One of the useful applications of the representation (8) is to investigate the role of the deterministic terms. Note that d_t cumulates in the process with a coefficient $C\Phi$, but that D_t does not, because $C\alpha\Upsilon = 0$. A leading special case is the model with $D_t = t$, and $d_t = 1$, which ensures that any linear combination of the components of x_t is allowed to have a linear trend. Note that if $D_t = t$ is not allowed in the model, that is $\Upsilon = 0$, then x_t has a trend given by $C\Phi t$, but the cointegration relation $\beta' x_t$ has no trend because $\beta' C\Phi = 0$.

2.3 Interpretation of cointegrating coefficients

Consider first a usual regression

$$x_{1t} = \gamma_2 x_{2t} + \gamma_3 x_{3t} + \varepsilon_t, \tag{9}$$

with i.i.d. errors ε_t which are independent of the processes x_{2t} and x_{3t}. The coefficient γ_2 is interpreted via a counterfactual experiment, that is, the coefficient γ_2 is the effect on x_{1t} of a change in x_{2t}, keeping x_{3t} constant.

The cointegration relations are long-run relations. This means that they have been there all the time, and they influence the movement of the process x_t via the adjustment coefficients α. The more the process $\beta' x_t$ deviates from $E\beta' x_t$, the more the adjustment coefficients pull the process back towards its mean. Another interpretation is that they are relations that would hold in the limit, provided all shocks in the model are set to zero after a time t.

It is therefore natural that interpretation of cointegration coefficients involves the notion of a long-run value. From the Granger Representation Theorem 8 applied to the model with no deterministic terms, it can be proved, see Johansen (2005), that

$$x_{\infty|t} = \lim_{h\to\infty} E(x_{t+h}|x_t, \ldots, x_{t-k+1}) = C(x_t - \sum_{i=1}^{k-1} \Gamma_i x_{t-i}) = C\sum_{i=1}^{t} \varepsilon_i + x_{\infty|0}.$$

This limiting conditional expectation is a long-run value of the process. Because $\beta' x_{\infty|t} = 0$, the point $x_{\infty|t}$ is in the attractor set $\{x : \beta' x = 0\} = sp\{\beta_\perp\}$, see Figure 1. Thus if the current value, x_t, is shifted to $x_t + h$, then

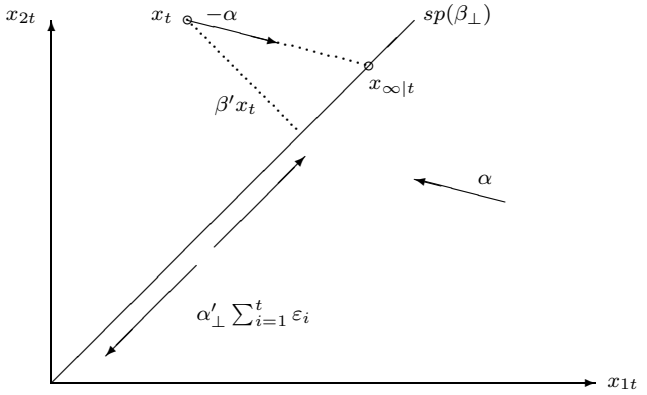

Fig. 1 In the model $\Delta x_t = \alpha\beta' x_{t-1} + \varepsilon_t$, the point $x_t = (x_{1t}, x_{2t})$ is moved towards the long-run value $x_{\infty|t}$ on the attractor set $\{x|\beta' x = 0\} = sp(\beta_\perp\}$ by the forces $-\alpha$ or $+\alpha$, and pushed along the attractor set by the common trends $\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i$.

the long-run value is shifted from $x_{\infty|t}$ to $x_{\infty|t} + Ch$, which is still a point in the attractor set because $\beta' x_{\infty|t} + \beta' Ch = 0$. If a given long-run change $k = C\xi$ in $x_{\infty|t}$ is needed, Γk is added to the current value x_t. This gives the long-run value

$$x_{\infty|t} + C\Gamma k = x_{\infty|t} + C\Gamma C\xi = x_{\infty|t} + C\xi = x_{\infty|t} + k,$$

where the identity $C\Gamma C = C$ is applied; see (7). This idea is now used to give an interpretation of a cointegration coefficient in the simple case of $r = 1$, $p = 3$, and where the relation is normalized on x_1

$$x_1 = \gamma_2 x_2 + \gamma_3 x_3, \tag{10}$$

so that $\beta' = (1, -\gamma_2, -\gamma_3)$. In order to give the usual interpretation as a regression coefficient (or elasticity if the measurements are in logs), a long-run change with the properties that x_2 changes by one, x_1 changes by γ_2, and x_3 is kept fixed, is needed. Thus the long-run change is $k = (\gamma_2, 1, 0)$, which satisfies $\beta' k = 0$, so that $k = C\xi$ for some ξ, and this can be achieved by moving the current value from x_t to $x_t + C\Gamma k$. In this sense, a coefficient in an identified cointegration relation can be interpreted as the effect of a long-run change to one variable on another, keeping all others fixed in the long run. More details can be found in Johansen (2005) and Proietti (1997).

3 Interpretation of the $I(1)$ Model for Cointegration

In this section model $H(r)$ defined by (2) is discussed. The parameters in $H(r)$ are

$$(\alpha, \beta, \Gamma_1, \ldots, \Gamma_{k-1}, \Upsilon, \Phi, \Omega).$$

All parameters vary freely and α and β are $p \times r$ matrices. The normalization and identification of α and β are discussed, and some examples of hypotheses on α and β, which are of economic interest are given.

3.1 The models $H(r)$

The models $H(r)$ are nested

$$H(0) \subset \cdots \subset H(r) \subset \cdots \subset H(p).$$

Here $H(p)$ is the unrestricted vector autoregressive model, so that α and β are unrestricted $p \times p$ matrices. The model $H(0)$ corresponds to the restriction $\alpha = \beta = 0$, which is the vector autoregressive model for the process in differences. Note that in order to have nested models, we allow in $H(r)$ for all processes with rank less than or equal to r.

The formulation allows us to derive likelihood ratio tests for the hypothesis $H(r)$ in the unrestricted model $H(p)$. These tests can be applied to check if one's prior knowledge of the number of cointegration relations is consistent with the data, or alternatively to construct an estimator of the cointegration rank.

Note that when the cointegration rank is r, the number of common trends is $p - r$. Thus if one can interpret the presence of r cointegration relations one should also interpret the presence of $p - r$ independent stochastic trends or $p - r$ driving forces in the data.

3.2 Normalization of parameters of the $I(1)$ model

The parameters α and β in (2) are not uniquely identified, because given any choice of α and β and any nonsingular $r \times r$ matrix ξ, the choice $\alpha\xi$ and $\beta\xi'^{-1}$ gives the same matrix $\Pi = \alpha\beta' = \alpha\xi(\beta\xi'^{-1})'$.

If $x_t = (x'_{1t}, x'_{2t})'$ and $\beta = (\beta'_1, \beta'_2)'$, with $|\beta_1| \neq 0$, we can solve the cointegration relations as

$$x_{1t} = \gamma' x_{2t} + u_t,$$

where u_t is stationary and $\gamma' = -(\beta'_1)^{-1}\beta'_2$. This represents cointegration as a regression equation. A normalization of this type is sometimes convenient for estimation and calculation of 'standard errors' of the estimate, see section 5.2, but many hypotheses are invariant with respect to a normalization of β, and thus, in a discussion of a test of such a hypothesis, β does not require normalization. As seen in subsection 3.3, many stable economic relations are expressed in terms of identifying restrictions, for which the regression formulation is not convenient.

From the Granger Representation Theorem we see that the $p - r$ common trends are the nonstationary random walks in $C\sum_{i=1}^{t} \varepsilon_i$, that is, can be chosen as $\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i$. For any full rank $(p-r)\times(p-r)$ matrix η, $\eta\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i$ could also be used as common trends because

$$C\sum_{i=1}^{t} \varepsilon_i = \beta_\perp(\alpha'_\perp \Gamma \beta_\perp)^{-1}(\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i) = \beta_\perp(\eta\alpha'_\perp \Gamma \beta_\perp)^{-1}(\eta\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i).$$

Thus identifying restrictions on the coefficients in $\alpha_\perp$ are needed to find their estimates and standard errors.

In the cointegration model there are therefore three different identification problems: one for the cointegration relations, one for the common trends, and finally one for the short run dynamics, if the model has simultaneous effects.

3.3 Hypotheses on long-run coefficients

The purpose of modeling economic data is to test hypotheses on the coefficients, thereby investigating whether the data supports an economic hypothesis or rejects it. In the example with the series $x_t = (p_t^{au}, p_t^{us}, i_t^{au}, i_t^{us}, e_t)'$

the hypothesis of PPP is formulated as the hypothesis that $(1, -1, 1, 0, 0)$ is a cointegration relation. Similarly, the hypothesis of price homogeneity is formulated as

$$R'\beta = (1, 1, 0, 0, 0)\beta = 0,$$

or equivalently as $\beta = R_\perp \varphi = H\varphi$, for some vector φ and $H = R_\perp$. The hypothesis that the interest rates are stationary is formulated as the hypothesis that the two vectors $(0, 0, 0, 1, 0)$ and $(0, 0, 0, 0, 1)$ are cointegration vectors. A general formulation of restrictions on each of r cointegration vectors, including a normalization, is

$$\beta = (h_1 + H_1\varphi_1, \ldots, h_r + H_r\varphi_r)\,. \tag{11}$$

Here h_i is $p \times 1$ and orthogonal to H_i which is $p \times (s_i - 1)$ of rank $s_i - 1$, so that $p - s_i$ restrictions are imposed on the vector β_i. Let $R_i = (h_i, H_i)_\perp$ then β_i satisfies the restrictions $R_i'\beta_i = 0$, and the normalization $(h_i'h_i)^{-1}h_i'\beta_i = 1$. Wald's identification criterion is that β_i is identified if

$$R_i'(\beta_1, \ldots, \beta_r) = r - 1.$$

3.4 Hypotheses on adjustment coefficients

The coefficients in α measure how the process adjusts to disequilibrium errors. The hypothesis of weak exogeneity is the hypothesis that some rows of α are zero; see Engle et al. (1983). The process x_t is decomposed as $x_t = (x_{1t}', x_{2t}')'$ and the matrices are decomposed similarly so that the model equations without deterministic terms become

$$\begin{aligned} \Delta x_{1t} &= \alpha_1\beta' x_{t-1} + \textstyle\sum_{i=1}^{k-1} \Gamma_{1i}\Delta x_{t-i} + \varepsilon_{1t}, \\ \Delta x_{2t} &= \alpha_2\beta' x_{t-1} + \textstyle\sum_{i=1}^{k-1} \Gamma_{2i}\Delta x_{t-i} + \varepsilon_{2t}. \end{aligned}$$

If $\alpha_2 = 0$, there is no levels feedback from $\beta' x_{t-1}$ to Δx_{2t}, and if the errors are Gaussian, x_{2t} is weakly exogenous for α_1, β. The conditional model for Δx_{1t} given Δx_{2t} and the past is

$$\Delta x_{1t} = \omega\Delta x_{2t} + \alpha_1\beta' x_{t-1} + \sum_{i=1}^{k-1}(\Gamma_{1i} - \omega\Gamma_{2i})\Delta x_{t-i} + \varepsilon_{1t} - \omega\varepsilon_{2t}, \tag{12}$$

where $\omega = \Omega_{12}\Omega_{22}^{-1}$. Thus full maximum likelihood inference on α_1 and β can be conducted in the conditional model (12).

An interpretation of the hypothesis of weak exogeneity is the following: if $\alpha_2 = 0$ then $\alpha_\perp$ contains the columns of $(0, I_{p-r})'$, so that $\sum_{i=1}^{t} \varepsilon_{2i}$ are common trends. Thus the errors in the equations for x_{2t} cumulate in the system and give rise to nonstationarity.

4 Likelihood Analysis of the $I(1)$ Model

This section contains first some comments on what aspects are important for checking for model misspecification, and then describes the calculation of reduced rank regression, introduced by Anderson (1951). Then reduced rank regression and modifications thereof are applied to estimate the parameters of the $I(1)$ model (2) and various submodels.

4.1 Checking the specifications of the model

In order to apply Gaussian maximum likelihood methods, the assumptions behind the model have to be checked carefully, so that one is convinced that the statistical model contains the density that describes the data. If this is not the case, the asymptotic results available from the Gaussian analysis need not hold. Methods for checking vector autoregressive models include choice of lag length, test for normality of residuals, tests for autocorrelation, and test for heteroscedasticity in errors. Asymptotic results for estimators and tests derived from the Gaussian likelihood turn out to be robust to some types of deviations from the above assumptions. Thus the limit results hold for i.i.d. errors with finite variance, and not just for Gaussian errors, but autocorrelated errors violate the asymptotic results, so autocorrelation has to be checked carefully.

Finally and perhaps most importantly, the assumption of constant parameters is crucial. In practice it is important to model outliers by suitable dummies, but it is also important to model breaks in the dynamics, breaks in the cointegration properties, breaks in the stationarity properties, etc. The papers by Seo (1998) and Hansen and Johansen (1999) contain some results on recursive tests in the cointegration model.

4.2 Reduced rank regression

Let u_t, w_t, and z_t be three multivariate time series of dimensions p_u, p_w, and p_z respectively. The algorithm of reduced rank regression, see Anderson (1951), can be described in the regression model

$$u_t = \alpha\beta' w_t + \Gamma z_t + \varepsilon_t, \tag{13}$$

where ε_t are the errors with variance Ω. The product moments are

$$S_{uw} = T^{-1} \sum_{t=1}^{T} u_t w_t',$$

and the residuals, which we get by regressing u_t on w_t, are

$$(u_t|w_t) = u_t - S_{uw}S_{ww}^{-1}w_t,$$

so that the conditional product moments are

$$S_{uw.z} = S_{uw} - S_{uz}S_{zz}^{-1}S_{zw} = T^{-1}\sum_{t=1}^{T}(u_t|z_t)(w_t|zv)',$$

$$S_{uu.w,z} = T^{-1}\sum_{t=1}^{T}(u_t|w_t,z_t)(u_t|w_t,z_t)' = S_{uu.w} - S_{uz.w}S_{zz.w}^{-1}S_{zu.w}.$$

Let $\Pi = \alpha\beta'$. The unrestricted regression estimates are

$$\widehat{\Pi} = S_{uw.z}S_{ww.z}^{-1},\ \widehat{\Gamma} = S_{uz.w}S_{zz.w}^{-1},\ \text{and}\ \widehat{\Omega} = S_{uu.w,z}.$$

Reduced rank regression of u_t on w_t corrected for z_t gives estimates of α, β and Ω in (13). First the eigenvalue problem

$$|\lambda S_{ww.z} - S_{wu.z}S_{uu.z}^{-1}S_{uw.z}| = 0 \tag{14}$$

is solved. The eigenvalues are ordered $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_{p_w}$, with corresponding eigenvectors $\widehat{v}_1, \ldots, \widehat{v}_{p_w}$. The reduced rank estimates of β, α, Γ, and Ω are given by

$$\begin{aligned}
&\widehat{\beta} = (\widehat{v}_1, \ldots, \widehat{v}_r),\\
&\widehat{\alpha} = S_{uw.z}\widehat{\beta},\\
&\widehat{\Gamma} = S_{uz.\widehat{\beta}'w}S_{zz.\widehat{\beta}'w}^{-1},\\
&\widehat{\Omega} = S_{uu.z} - S_{uw.z}\widehat{\beta}\widehat{\beta}'S_{wu.z},\\
&|\widehat{\Omega}| = |S_{uu.z}|\prod_{i=1}^{r}(1-\widehat{\lambda}_i).
\end{aligned} \tag{15}$$

The eigenvectors are orthogonal because $\widehat{v}_i'S_{ww.z}\widehat{v}_j = 0$ for $i \neq j$, and are normalized by $\widehat{v}_i'S_{ww.z}\widehat{v}_i = 1$. The calculations described here are called a reduced rank regression and are denoted by $RRR(u_t, w_t|z_t)$.

4.3 Maximum likelihood estimation in the $I(1)$ model and derivation of the rank test

Consider the $I(1)$ model given by equation (2). Note that the multiplier $\alpha\Upsilon$ of D_t is restricted to be proportional to α so that, by the Granger Representation Theorem, D_t does not cumulate in the process. It is assumed for the derivations of maximum likelihood estimators and likelihood ratio tests that ε_t is i.i.d. $N_p(0, \Omega)$, but for asymptotic results the Gaussian assumption is not needed. The Gaussian likelihood function shows that the maximum

likelihood estimator can be found by the reduced rank regression

$$RRR(\Delta x_t, (x'_{t-1}, D'_t)' | \Delta x_{t-1}, \ldots, \Delta x_{t-k+1}, d_t).$$

It is convenient to introduce the notation for residuals

$$\begin{aligned} R_{0t} &= (\Delta x_t | \Delta x_{t-1}, \ldots, \Delta x_{t-k+1}, d_t) \\ R_{1t} &= ((x'_{t-1}, D'_t)' | \Delta x_{t-1}, \ldots, \Delta x_{t-k+1}, d_t) \end{aligned}$$

and product moments

$$S_{ij} = T^{-1} \sum_{t=1}^{T} R_{it} R'_{jt}.$$

The estimates are given by (15), and the maximized likelihood is, apart from a constant, given by

$$L_{\max}^{-2/T} = |\widehat{\Omega}| = |S_{00}| \prod_{i=1}^{r} (1 - \widehat{\lambda}_i). \tag{16}$$

Note that all the models $H(r)$, $r = 0, \ldots, p$, have been solved by the same eigenvalue calculation. The maximized likelihood is given for each r by (16) and by dividing the maximized likelihood function for r with the corresponding expression for $r = p$, the likelihood ratio test for cointegration rank is obtained:

$$-2\log LR(H(r)|H(p)) = -T \sum_{i=r+1}^{p} \log(1 - \widehat{\lambda}_i). \tag{17}$$

This statistic was considered by Bartlett (1948) for testing canonical correlations. The asymptotic distribution of this test statistic and the estimators are discussed in section 5.

The model obtained under the hypothesis $\beta = H\varphi$, is analyzed by

$$RRR(\Delta x_t, (H' x'_{t-1}, D'_t)' | \Delta x_{t-1}, \ldots, \Delta x_{t-k+1}, d_t),$$

and a number of hypotheses of this type for β and α can be solved in the same way, but the more general hypothesis

$$\beta = (h_1 + H_1\varphi_1, \ldots, h_r + H_r\varphi_r),$$

cannot be solved by reduced rank regression. With $\alpha = (\alpha_1, \ldots, \alpha_r)$ and $\Upsilon = (\Upsilon'_1, \ldots, \Upsilon'_r)'$, equation (2) becomes

$$\Delta x_t = \sum_{j=1}^{r} \alpha_j ((h_j + H_j\varphi_j)' x_{t-1} + \Upsilon_j D_t) + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \Phi d_t + \varepsilon_t.$$

This is reduced rank regression, but there are r reduced rank matrices $\alpha_j(1, \varphi_j', \Upsilon_j)$ of rank one. The solution is not given by an eigenvalue problem, but there is a simple modification of the reduced rank algorithm, which is easy to implement and is quite often found to converge. The algorithm has the property that the likelihood function is maximized in each step. The algorithm switches between reduced rank regressions of Δx_t on $(x_{t-1}'(H_i, h_i), D_t')'$ corrected for

$$(((h_j + H_j\varphi_j)'x_{t-1} + \Upsilon_j D_t)_{j\neq i}, \Delta x_{t-1}, \ldots, \Delta x_{t-k+1}, d_t).$$

This result can immediately be applied to calculate likelihood ratio tests for many different restrictions on the coefficients of the cointegration relations. Thus, in particular, this can give a test of over-identifying restrictions.

5 Asymptotic Analysis

A discussion of the most important aspects of the asymptotic analysis of the cointegration model is given. This includes the result that the rank test requires a family of Dickey-Fuller type distributions, depending on the specification of the deterministic terms of the model. The asymptotic distribution of $\widehat{\beta}$ is mixed Gaussian and that of the remaining parameters is Gaussian, so that tests for hypotheses on the parameters are asymptotically distributed as χ^2. All results are taken from Johansen (1996).

5.1 Asymptotic distribution of the rank test

The asymptotic distribution of the rank test is given in case the process has a linear trend.

Theorem 3 *Let ε_t be i.i.d. $(0, \Omega)$ and assume that $D_t = t$ and $d_t = 1$, in model* (2)*. Under the assumptions that the cointegration rank is r, the asymptotic distribution of the likelihood ratio test statistic* (17) *is*

$$-2logLR(H(r)|H(p)) \xrightarrow{d} tr\{\int_0^1 (dB)F'(\int_0^1 FF'du)^{-1}\int_0^1 F(dB)'\}, \quad (18)$$

where F is defined by

$$F(u) = \left(\begin{array}{c|c} B(u) & \\ u & 1 \end{array}\right),$$

and $B(u)$ is the $p-r$ dimensional standard Brownian motion.

The limit distribution is tabulated by simulating the distribution of the test of no cointegration in the model for a $p-r$ dimensional model with one lag and the same deterministic terms. Note that the limit distribution does not depend on the parameters $(\Gamma_1, \ldots, \Gamma_{k-1}, \Upsilon, \Phi, \Omega)$, but only on $p-r$, the number of common trends, and the presence of the linear trend. For finite samples, however, the dependence on the parameters can be quite pronounced. A small sample correction for the test has been given in Johansen (2002), and the bootstrap has been investigated by Swensen (2006).

In the model without deterministics the same result holds, but with $F(u) = B(u)$. A special case of this, for $p = 1$, is the Dickey-Fuller test and the distributions (18) are called the Dickey–Fuller distributions with $p-r$ degrees of freedom; see Dickey and Fuller (1981).

The asymptotic distribution of the test statistic for rank depends on the deterministic terms in the model. It follows from the Granger Representation Theorem that the deterministic term d_t is cumulated to $C\Phi\sum_{i=1}^{t} d_i$. In deriving the asymptotics, x_t is normalized by $T^{-1/2}$. If $\sum_{i=1}^{t} d_i$ is bounded, this normalization implies that the limit distribution does not depend on the precise form of $\sum_{i=1}^{t} d_i$. Thus, if d_t is a centered seasonal dummy, or an 'innovation dummy' $d_t = 1_{\{t=t_0\}}$, it does not change the asymptotic distribution. If, on the other hand, a 'step dummy' $d_t = 1_{\{t\geq t_0\}}$ is included, then the cumulation of this is a broken linear trend, and that influences the limit distribution and requires special tables; see Johansen et al. (2006d).

5.2 Asymptotic distribution of the estimators

The main result here is that the estimator of β, suitably normalized, converges to a mixed Gaussian distribution, even when estimated under continuously differentiable restrictions, see Johansen (1991). This result implies that likelihood ratio tests on β are asymptotically χ^2 distributed. Furthermore the estimators of the adjustment parameters α and the short-run parameters Γ_i are asymptotically Gaussian and asymptotically independent of the estimator for β.

In order to illustrate these results, the asymptotic distribution of $\widehat{\beta}$ for $r = 2$ is given, when β is identified by

$$\beta = (h_1 + H_1\varphi_1, h_2 + H_2\varphi_2). \tag{19}$$

Theorem 4 *In model* (2) *without deterministic terms and* ε_t *i.i.d.* $(0, \Omega)$, *the asymptotic distribution of* $Tvec(\widehat{\beta} - \beta)$ *is given by*

$$\begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix} \begin{pmatrix} \rho_{11}H_1'\mathcal{G}H_1 & \rho_{12}H_1'\mathcal{G}H_2 \\ \rho_{21}H_2'\mathcal{G}H_1 & \rho_{22}H_2'\mathcal{G}H_2 \end{pmatrix}^{-1} \begin{pmatrix} H_1' \int_0^1 G(dV_1) \\ H_2' \int_0^1 G(dV_2) \end{pmatrix}, \tag{20}$$

where

$$T^{-1/2}x_{[Tu]} \xrightarrow{d} G = CW,$$
$$T^{-1}S_{11} \xrightarrow{d} \mathcal{G} = C\int_0^1 WW'duC',$$

and

$$V = \alpha'\Omega^{-1}W = (V_1, V_2)',$$
$$\rho_{ij} = \alpha_i'\Omega^{-1}\alpha_j.$$

The estimators of the remaining parameters are asymptotically Gaussian and asymptotically independent of $\widehat{\beta}$.

Note that G and V are independent Brownian motions so that the limit distribution is mixed Gaussian and the asymptotic conditional distribution given G is Gaussian with asymptotic conditional variance

$$\begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix} \begin{pmatrix} \rho_{11}H_1'\mathcal{G}H_1 & \rho_{12}H_1'\mathcal{G}H_2 \\ \rho_{21}H_2'\mathcal{G}H_1 & \rho_{22}H_2'\mathcal{G}H_2 \end{pmatrix}^{-1} \begin{pmatrix} H_1' & 0 \\ 0 & H_2' \end{pmatrix}.$$

A consistent estimator for the asymptotic conditional variance is

$$T\begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix} \begin{pmatrix} \widehat{\rho}_{11}H_1'S_{11}H_1 & \widehat{\rho}_{12}H_1'S_{11}H_2 \\ \widehat{\rho}_{21}H_1'S_{11}H_2 & \widehat{\rho}_{22}H_2'S_{11}H_2 \end{pmatrix}^{-1} \begin{pmatrix} H_1' & 0 \\ 0 & H_2' \end{pmatrix}. \tag{21}$$

In order to interpret these results, note that the observed information about β in the data (keeping other parameters fixed) is given by

$$\mathcal{J}_T = T\begin{pmatrix} \rho_{11}H_1'S_{11}H_1 & \rho_{12}H_1'S_{11}H_2 \\ \rho_{21}H_2'S_{11}H_1 & \rho_{22}H_2'S_{11}H_2 \end{pmatrix},$$

which normalized by T^2 converges to the stochastic limit

$$\mathcal{J} = \begin{pmatrix} \rho_{11}H_1'\mathcal{G}H_1 & \rho_{12}H_1'\mathcal{G}H_2 \\ \rho_{21}H_2'\mathcal{G}H_1 & \rho_{22}H_2'\mathcal{G}H_2 \end{pmatrix}.$$

Thus the result (20) states that, given the asymptotic information or equivalently the limit of the common trends, $\alpha_\perp'W$, the limit distribution of $T(\widehat{\beta}-\beta)$ is Gaussian with a variance that is a function of the inverse limit information. Hence the asymptotic distribution of

$$\mathcal{J}_T^{1/2}\begin{pmatrix} \bar{H}_1'(\widehat{\beta}_1 - \beta_1) \\ \bar{H}_2'(\widehat{\beta}_2 - \beta_2) \end{pmatrix}$$

is a standard Gaussian distribution. Here $\bar{H}_i' = (H_i'H_i)^{-1}H_i'$. This implies that Wald and therefore likelihood ratio tests on β can be conducted using the asymptotic χ^2 distribution.

It is therefore possible to scale the deviations $\widehat{\beta} - \beta$ in order to obtain an asymptotic Gaussian distribution. Note that the scaling matrix $\mathcal{J}_T^{1/2}$ is not an

estimate of the asymptotic variance of $\widehat{\beta}$, but an estimate of the asymptotic *conditional* variance given the information in the data. It is therefore not the asymptotic distribution of $\widehat{\beta}$ that is used for inference, but the *conditional* distribution given the information; see Basawa and Scott (1983) or Johansen (1995) for a discussion. Finally the result on the likelihood ratio test for the restrictions given in (19) is formulated.

Theorem 5 *Let ε_t be i.i.d. $(0, \Omega)$. The asymptotic distribution of the likelihood ratio test statistic for the restrictions* (19) *in model* (2) *with no deterministic terms is χ^2 with degrees of freedom given by $\sum_{i=1}^{r}(p - r - s_i + 1)$.*

This result is taken from Johansen (1996), and a small sample correction for some tests on β has been developed in Johansen (2000).

6 Further Topics in the Area of Cointegration

It is mentioned here how the $I(1)$ model can be applied to test hypotheses implied by rational expectations. The basic model for $I(1)$ processes can be extended to other models of nonstationarity. In particular models for seasonal roots, explosive roots, $I(2)$ processes, fractionally integrated processes and nonlinear cointegration. We discuss here models for $I(2)$ processes, and refer to the paper by Lange and Rahbeck (2008) for some models of nonlinear cointegration.

6.1 Rational expectations

Many economic models operate with the concept of rational or model based expectations; see Hansen and Sargent (1991). An example of such a formulation is uncovered interest parity,

$$\Delta^e e_{t+1} = i_t^1 - i_t^2, \tag{22}$$

which expresses a balance between interest rates in two countries and economic expectations of exchange rate changes. If a vector autoregressive model

$$\Delta x_t = \alpha\beta' x_{t-1} + \Gamma_1 \Delta x_{t-1} + \varepsilon_t, \tag{23}$$

fits the data $x_t = (e_t, i_t^1, i_t^2)'$, the assumption of model based expectations, Muth (1961), means that $\Delta^e e_{t+1}$ can be replaced by the conditional expectation $E_t \Delta e_{t+1}$ based upon model (23). That is,

$$\Delta^e e_{t+1} = E_t \Delta e_{t+1} = \alpha_1 \beta' x_t + \Gamma_{11} \Delta x_t.$$

Assumption (22) implies the identity

$$i_t^1 - i_t^2 = \alpha_1 \beta' x_t + \Gamma_{11} \Delta x_t.$$

Hence the cointegration relation is

$$\beta' x_t = i_t^1 - i_t^2,$$

and the other parameters are restricted by $\alpha_1 = 1$, and $\Gamma_{11} = 0$. Thus, the hypothesis (22) implies a number of testable restrictions on the vector autoregressive model. The implications of model based expectations for the cointegrated vector autoregressive model is explored in Johansen and Swensen (2004), where it is shown that, as in the example above, rational expectation restrictions assume testable information on cointegration relations and short-run adjustments. It is demonstrated how estimation under rational expectation restrictions can be performed by regression and reduced rank regression in certain cases.

6.2 The $I(2)$ model

It is sometimes found that inflation rates are best described by $I(1)$ processes and then log prices are $I(2)$. In such a case $\alpha_\perp' \Gamma \beta_\perp$ has reduced rank; see (5). Under this condition model (2) can be parametrized as

$$\Delta^2 x_t = \alpha(\beta' x_{t-1} + \psi' \Delta x_{t-1}) + \Omega \alpha_\perp (\alpha_\perp' \Omega \alpha_\perp)^{-1} \kappa' \tau' \Delta x_{t-1} + \varepsilon_t, \qquad (24)$$

where α and β are $p \times r$ and τ is $p \times (r+s)$, or equivalently as

$$\Delta^2 x_t = \alpha \begin{pmatrix} \beta \\ \delta' \end{pmatrix}' \begin{pmatrix} x_{t-1} \\ \bar{\tau}_\perp' \Delta x_{t-1} \end{pmatrix} + \zeta \tau' \Delta x_{t-1} + \varepsilon_t, \qquad (25)$$

where

$$\delta = \psi' \tau_\perp, \ \zeta = \alpha \psi' \bar{\tau} + \Omega \alpha_\perp (\alpha_\perp' \Omega \alpha_\perp)^{-1} \kappa';$$

see Johansen (1997) and Paruolo and Rahbek (1999). Under suitable conditions on the parameters, the solution of equations (24) or (25) has the form

$$x_t = C_2 \sum_{i=1}^{t} \sum_{j=1}^{i} \varepsilon_j + C_1 \sum_{i=1}^{t} \varepsilon_i + A_1 + t A_2 + y_t,$$

where y_t is stationary and C_1 and C_2 are functions of the model parameters. One can prove that the processes $\Delta^2 x_t$, $\beta' x_t + \psi' \Delta x_t$, and $\tau' \Delta x_t$ are stationary. Thus $\tau' x_t$ are cointegration relations from $I(2)$ to $I(1)$. The model also allows for multicointegration, that is, cointegration between levels and

differences because $\beta' x_t + \psi' \Delta x_t$ is stationary; see Engle and Yoo (1991). Maximum likelihood estimation can be performed by a switching algorithm using the two parametrizations given in (24) and (25). The same techniques can be used for a number of hypotheses on the cointegration parameters β and τ.

The asymptotic theory of likelihood ratio tests and maximum likelihood estimators is developed by Johansen (1997, 2006c), Rahbek et al. (1999), Paruolo (1996, 2000), Boswijk (2000) and Nielsen and Rahbek (2004). It is shown that the likelihood ratio test for rank involves not only Brownian motion, but also integrated Brownian motion and hence some new Dickey-Fuller type distributions that have to be simulated. The asymptotic distribution of the maximum likelihood estimator is quite involved, as it is not mixed Gaussian, but many hypotheses still allow asymptotic χ^2 inference; see Johansen (2006c).

7 Concluding Remarks

What has been developed for the cointegrated vector autoregressive model is a set of useful tools for the analysis of macroeconomic and financial time series. The theory is part of many textbooks, and software for the analysis of data has been implemented in several packages, e.g. in CATS in RATS, Givewin, Eviews, Microfit, Shazam, R, etc.

Many theoretical problems remain unsolved, however. We mention here three important problems for future development.

1. The analysis of models for time series strongly relies on asymptotic methods, and it is often a problem to obtain sufficiently long series in economics which actually measure the same variables for the whole period. Therefore periods which can be modelled by constant parameters are often rather short, and it is therefore extremely important to develop methods for small sample correction of the asymptotic results. Such methods can be analytic or simulation based. When these will become part of the software packages, and are routinely applied, they will ensure more reliable inference.

2. A very interesting and promising development lies in the analysis of cointegration in nonlinear time series, where the statistical theory is still in its beginning. Many different types of nonlinearities are possible, and the theory has to be developed in close contact with applications in order to ensure that useful models and concepts are developed; see the overview Lange and Rahbeck (2008).

3. Most importantly, however, is the development of an economic theory which takes into account the findings of empirical analyses of nonstationary economic data. For a long time, regression analysis and correlations have been standard tools for quantitative analysis of relations between variables in economics. Economic theory has incorporated these techniques in order to

learn from data. In the same way economic theory should be developed to incorporate nonstationarity of data and develop theories consistent with the findings of empirical cointegration analyses.

References

Ahn, S.K. and Reinsel, G.C. (1990): Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association* **85**, 813–823.

Anderson, T.W. (1951): Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* **22**, 327–351.

Anderson, T.W. (1971): *The statistical analysis of time series.* Wiley, New York.

Banerjee, A., Dolado J.J., Galbraith J.W. and Hendry D.F.(1993): *Co-integration error-correction and the econometric analysis of nonstationary data.* Oxford University Press, Oxford.

Bartlett, M. (1948): A note on the statistical estimation of the demand and supply relations from time series. *Econometrica* **16**, 323–329.

Basawa, I.V. and Scott, D.J. (1983): *Asymptotic optimal inference for non-ergodic models.* Springer, New York.

Boswijk, P. (2000): Mixed normality and ancillarity in $I(2)$ systems. *Econometric Theory* **16**, 878–904.

Campbell, J. and Shiller, R.J. (1987): Cointegration and tests of present value models. *Journal of Political Economy* **95**, 1062–1088.

Dickey, D.A. and Fuller, W.A. (1981): Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49**, 1057–1072.

Engle, R.F. and Granger, C.W.J. (1987): Co-integration and error correction: Representation, estimation and testing. *Econometrica* **55**, 251–276.

Engle, R.F., Hendry, D.F. and Richard. J.-F. (1983): Exogeneity. *Econometrica* **51**, 277–304.

Engle, R.F. and Yoo, B.S. (1991): Cointegrated economic time series: A survey with new results. *In: Granger, C.W.J. and Engle, R.F. (Eds.): Long-run economic relations. Readings in cointegration.* Oxford University Press, Oxford.

Granger, C.W.J. (1983): Cointegrated variables and error correction models. *UCSD Discussion paper* **83–13a**.

Hamilton, J.D. (1994): *Time series analysis.* Princeton University Press, Princeton New Jersey.

Hansen. L.P. and Sargent, T.J. (1991): Exact linear rational expectations models: Specification and estimation. *In: Hansen, L.P. and Sargent, T.J. (Eds.): Rational expectations econometrics.* Westview Press, Boulder.

Hansen, H. and Johansen, S.(1999): Some tests for parameter constancy in the cointegrated VAR. *The Econometrics Journal* **2**, 306–333.

Hendry, D.F. (1995): *Dynamic econometrics.* Oxford University Press, Oxford.

Johansen, S. (1988): Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**, 231–254.

Johansen, S. (1995): The role of ancillarity in inference for nonstationary variables. *Economic Journal* **13**, 302–320.

Johansen, S. (1991): Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551–1580.

Johansen, S. (1996): Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press, Oxford.

Johansen, S. (1997): Likelihood analysis of the $I(2)$ model. *Scandinavian Journal of Statistics* **24**, 433–462.

Johansen, S. (2000): A Bartlett correction factor for tests on the cointegration relations. *Econometric Theory* **16**, 740–778.

Johansen, S. (2002): A small sample correction of the test for cointegration rank in the vector autoregressive model. *Econometrica* **70**, 1929–1961.

Johansen, S. (2005): The interpretation of cointegration coefficients in the cointegrated vector autoregressive model. *Oxford Bulletin of Economics and Statistics* **67**, 93–104.

Johansen, S. (2006a): Cointegration: a survey. *In: Mills, T.C. and Patterson, K. (Eds.): Palgrave handbook of econometrics: Volume 1, Econometric theory*. Palgrave Macmillan, Basingstoke.

Johansen, S. (2006b): Representation of cointegrated autoregressive processes with application to fractional processes. *Forthcoming in Econometric Reviews*.

Johansen, S. (2006c): Statistical analysis of hypotheses on the cointegration relations in the $I(2)$ model. *Journal of Econometrics* **132**, 81–115.

Johansen, S., Mosconi, R. and Nielsen, B. (2000): Cointegration analysis in the presence of structural breaks in the deterministic trend. *The Econometrics Journal* **3**, 1–34.

Johansen, S. and Swensen, A.R. (2004): More on testing exact rational expectations in vector autoregressive models: Restricted drift term. *The Econometrics Journal* **7**, 389–397.

Juselius, K. and MacDonald, R. (2004): The international parities between USA and Japan. *Japan and the World Economy* **16**, 17–34.

Juselius, K. (2006): *The cointegrated VAR model: Econometric methodology and macroeconomic applications*. Oxford University Press, Oxford.

Lange, T. and Rahbek, A. (2008): An introduction to regime switching time series models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 871–887. Springer, New York.

Lütkepohl, H. (2006): Introduction to multiple times series analysis. Springer, New York.

Muth, J.F. (1961): Rational expectations and the theory of price movements. *Econometrica* **29**, 315–335.

Nielsen, H.B. and Rahbek, A. (2004): Likelihood ratio testing for cointegration ranks in $I(2)$ Models. *Forthcoming Econometric Theory*.

Nyblom, J. and Harvey, A. (2000): Tests of common stochastic trends. *Econometric Theory* **16**, 176–199.

Paruolo, P. (1996): On the determination of integration indices in $I(2)$ systems. *Journal of Econometrics* **72**, 313–356.

Paruolo, P. and Rahbek, A. (1999): Weak exogeneity in $I(2)$ VAR systems. *Journal of Econometrics* **93**, 281–308.

Paruolo, P. (2000): Asymptotic efficiency of the two stage estimator in $I(2)$ systems. *Econometric Theory* **16**, 524–550.

Phillips, P.C.B. (1991): Optimal inference in cointegrated systems. *Econometrica* **59**, 283–306.

Phillips, P.C.B. and Hansen, B.E. (1990): Statistical inference on instrumental variables regression with $I(1)$ processes. *Review of Economic Studies* **57**, 99–124.

Proietti, T. (1997): Short-run dynamics in cointegrated systems. *Oxford Bulletin of Economics and Statistics* **59**, 405–422.

Rahbek, A., Kongsted, H.C. and Jørgensen, C. (1999): Trend-stationarity in the $I(2)$ cointegration model. *Journal of Econometrics* **90**, 265–289.

Seo, B. (1998): Tests for structural change in cointegrated systems. *Econometric Theory* **14**, 222–259.

Stock, J.H. (1987): Asymptotic properties of least squares estimates of cointegration vectors. *Econometrica* **55**, 1035–1056.

Swensen, A.R. (2006): Bootstrap algorithms for testing and determining the cointegrating rank in VAR models. *Forthcoming Econometric Theory*.

Watson, M. (1994): Vector autoregressions and cointegration. *In: Engle, R.F. and McFadden, D. (Eds.): Handbook of econometrics Vol. 4*. North Holland Publishing Company, The Netherlands.

Time Series with Roots on or Near the Unit Circle

Ngai Hang Chan*

Abstract This paper reviews some of the developments of the unit root and near unit root time series. It gives an overview of this important topic and describes the impact of some of the recent progress on subsequent research.

1 Introduction

The field of unit root time series has received considerable attention in both the statistics and the econometric literature during the last 30 years. Research under the umbrellas of unit root, near unit root, nonstationary, nearly nonstationary, integrated and near-integrated processes has been pursued actively. In essence, all these titles referred to time series with autoregressive (AR) roots on or near the unit circle. Since the seminal paper of White (1958), numerous attempts have been devoted to studying the asymptotic behavior of the least-squares estimate (LSE) of the AR coefficient of a unit root AR model. Since several review articles have been written on this subject by econometricians, we do not attempt to offer a comprehensive review of this topic. In this paper, we focus on some of the strategic developments in this area related to statistics and offer some future directions. Succinct reviews on early developments on this subject can be found in Fuller (1996) and Chan (2002) and the references therein.

This paper is organized as follows. In Sections 2.1–2.3, we review the developments of the unit root problem according to estimation, inference and

Ngai Hang Chan
Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong, e-mail: nhchan@sta.cuhk.edu.hk

* The author would like to thank Richard Davis, one of the editors, for helpful comments. This research was supported in part by HKSAR-RGC grants CUHK400305 and CUHK400306.

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_30,

model selection for AR(1) and AR(p) models. Here, we also provide a road map for certain portions of the extensive literature on these topics, and the impact of some of the seminal work in related areas. Section 3 contains ideas on new developments and some concluding remarks.

2 Unit Root Models

Developments of the unit root problem can be classified into three stages: estimation, inference and model selection. Throughout these stages, there are some underlying common themes. The seminal paper of Lai et al. (1978) considered the strong consistency property of the least squares estimate (LSE) of a multiple regression model. Specifically, consider the model

$$y_i = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i + \epsilon_i, \ i = 1, 2, \ldots, \tag{1}$$

where the ϵ_i are unobservable errors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are unknown parameters and y_i is the observable response corresponding to the design vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Then

$$\boldsymbol{b}_n = (\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}})^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i y_i \tag{2}$$

is the LSE of the unknown parameter vector $\boldsymbol{\beta}$ based on the observations $\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_n, y_n$. Herein, the unobservable sequence $\{\epsilon_n\}$ is assumed to be a martingale difference sequence with respect to an increasing sequence of sigma fields $\mathcal{F}_n$ satisfying a Lyapunov condition

$$\sup_n E(|\epsilon_n|^{\gamma} | \mathcal{F}_{n-1}) < \infty \text{ almost surely (a.s.) for some } \gamma > 2. \tag{3}$$

By assuming that the design vector at stage n is adaptable, i.e., $\boldsymbol{x}_n$ is $\mathcal{F}_{n-1}$ measurable, Lai and Wei (1982a) proved the following.

Theorem 1 $\boldsymbol{b}_n \to \boldsymbol{\beta}$ *a.s. if*

$$\lambda_{\min}(n) \to \infty \ \textit{a.s. and} \ \log \lambda_{\max}(n) = o(\lambda_{\min}(n)) \ \textit{a.s.}, \tag{4}$$

where $\lambda_{\min}(n)$ *and* $\lambda_{\max}(n)$ *denote respectively the minimum and the maximum eigenvalues of the design matrix* $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$ *at stage* n.

This result extended the earlier work of Anderson and Taylor (1979) and constituted the turning point of the study of strong consistency in the regression literature. On the basis of this groundbreaking result, Lai and Wei (1982b, 1985) established the strong consistency of the LSE for a general AR model, irrespective of the location of its characteristic roots. This is one of the most general results concerning the strong consistency of the LSE of a general

nonstationary AR(p) model. Tiao and Tsay (1983) subsequently considered the weak consistency property of the LSE for a nonstationary AR moving average model. Recently, Nielsen (2005) extended Lai and Wei's result to the general vector AR (VAR) model with deterministic components.

2.1 First order

Back in 1958, White (1958), among others, first showed that the LSE of the autoregressive coefficient of a nonstationary AR(1) model, i.e., when the autoregressive coefficient equals 1, converges in distribution to a functional of a stochastic integral of a standard Brownian motion. It turns out that this model has a strong bearing in the econometric literature in testing whether or not a time series is a random walk, the so-called unit root testing problem. After Dickey and Fuller (1979) established the form of this limiting distribution as a ratio of sums of independent and identically distributed random variables, the unit root testing problem became a topical issue in econometrics. Numerous articles were written on this topic and two elementary surveys on the statistical and econometric literature were given in Dickey et al. (1986) and Stock and Watson (1987), respectively.

With the strong consistency results, the next natural question is asymptotic inference. One is interested in the limiting distributions of the LSEs of the parameters of a general nonstationary autoregressive model. Chan and Wei (1987) considered the AR(1) model when the autoregressive coefficient converges to 1 asymptotically. Instead of a Brownian motion, Chan and Wei (1987) showed that the limiting distribution of the LSE of a nearly nonstationary AR(1) model converges weakly to a functional of an Ornstein–Uhlenbeck process. Specifically, consider a triangular array of first-order AR processes

$$y_{t,n} = \beta_n y_{t-1,n} + \epsilon_t, \quad t = 1, \ldots, n, \tag{5}$$

where $\beta_n = 1 - \gamma/n$, γ is a real number, $y_{0,n} = 0$ for all n and $\{\epsilon_t\}$ is a martingale difference sequence satisfying (3). This is known as the nearly nonstationary or near-integrated (Phillips (1987)) time series. Note that when $\gamma = 0$, (5) reduces to the traditional unit root model.

Theorem 2 *Let the time series $\{y_{t,n}\}$ follow (5) with the innovation sequence $\{\epsilon_t\}$ satisfying (3). Let the LSE of β_n be $b_n = (\sum_{t=1}^n y_{t-1,n}\epsilon_t)/(\sum_{t=1}^n y_{t-1,n}^2)$. Then as $n \to \infty$,*

$$n(b_n - \beta_n) \to_D \mathcal{L}(\gamma) := \frac{\int_0^1 X(t)\, dX(t)}{\int_0^1 X^2(t)\, dt},$$

where $\to_D$ denotes convergence in distribution and $X(t)$ is the Ornstein–Uhlenbeck process satisfying the diffusion equation

$$dX(t) = -\gamma X(t)\,dt + dW(t),$$

where $X(0) = 0$ *and* $W(t)$ *is a standard Brownian motion.*

This particular result encompasses the unit root case of White (1958) when $\gamma = 0$. In this case, the autoregressive coefficient $\beta_n = 1$ and $\mathcal{L}(0) = \int_0^1 W(t)\,dW(t) / \int_0^1 W^2(t)\,dt$, which is the limiting distribution of the Dickey–Fuller statistic . Using reproducing kernels, Chan (1988) further developed this limiting form as sums of iid random variables. The near-integrated notion was formulated with reference to the work of LeCam concerning limiting experiments in terms of contiguous alternatives. This idea was later explored by Jeganathan (1991, 1995), who generalized the near-integrated notion to a general AR(p) case and introduced the idea of local asymptotic Brownian functional in studying optimality issues. In a spectral setting, Dahlhaus (1985) considered both tapered and nontapered Yule–Walker estimates for near-integrated models. Since then, numerous extensions have been carried out by econometricians and statisticians. For example, on the statistical front, Cox and Llatas (1991) considered the M-estimation of a near nonstationary process, Pham (1992) and Ing (2001) studied the bias and prediction mean square error expansion of the LSE of a near unit root model, Basawa et al. (1991) investigated the bootstrap estimate of a unit root model, Larsson (1998) considered the Barlett correction property of the unit root test statistics and Chan (1988) extended the notion of near unit root to a seasonal model. On the econometric front, the issues of testing for the unit root hypothesis for econometric series was considered in Phillips and Perron (1988) and Perron (1989) among others. Many of these results were explored further by various people under the topics of trend breaks and cointegration. Interested readers can consult the review articles of Stock (1994) and Watson (1994) on these topics. Finally, it should also be pointed out that while extensive studies were pursued by statisticians and econometricians alike, one of the earlier developments in this problem was given in the monograph of Arató (1982) and later in an unpublished thesis of Bobkoski (1983).

A problem closely related to the unit root AR(1) model is a unit root MA(1) model given by

$$y_t = \epsilon_t - \theta\epsilon_{t-1}, \tag{6}$$

with $\{\epsilon_t\}$ satisfying (3). Aysmptotic properties of the maximum likelihood estimate $\widehat{\theta}_n$ of the MA coefficient when $\theta = 1$ constitute an actively pursued area. Cryer and Ledolter (1981) first examined the consistency property of $\widehat{\theta}_n$. Davis and Dunsmuir (1996) derived the limiting distribution of the maximum likelihood estimator of θ under the unit root $\theta = 1$ and $\theta = 1 - c/n$ (for a constant $c > 0$) setups. Interestingly, there is a significant pileup probability, i.e., $P(\widehat{\theta_n} = 1)$ has a nonzero limit for all values of $c > 0$. The limiting pileup probabilities can be quite large, especially for small values of c. Saikkonen and Luukkonen (1993) examined the testing for a moving average unit root issue. Using the idea of the so-called derived process, Chan and Tsay (1996)

studied the limiting distribution of the maximum likelihood estimate for a general unit root MA(q) model. A systematic account on the unit root moving average model is given in Chapter 8 of Tanaka (1996).

2.2 AR(p) models

Chan and Wei (1988) considered the limiting distributions of the LSE of a general nonstationary AR(p) model when the characteristic roots lie on or outside the unit circle, each of which may have different multiplicities. This was the first comprehensive treatment of the LSE for a general nonstationary AR(p) model, and it was shown in Chan and Wei (1988) that the locations of the roots of the time series played an important role in characterizing the limiting distributions. Specifically, they considered a general nonstationary AR(p) model

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \epsilon_t. \tag{7}$$

In (7), the AR polynomial $\beta(z) = 1 - \beta_1 z - \cdots - \beta_p z^p$ has roots lying on or outside the unit circle. That is,

$$\beta(z) = (1-z)^a (1+z)^b \Pi_{k=1}^{\ell} (1 - 2\cos\theta_k z + z^2)^{d_k} \psi(z), \tag{8}$$

where a, b, ℓ and d_k are nonnegative integers, θ_k belongs to $(0, \pi)$ and $\psi(z)$ is a polynomial of order $r = p - (a + b + 2d_1 + \cdots + 2d_k)$ that has all roots outside the unit disk.

When the underlying model is stationary with all roots lying outside the unit circle, classical central limit theorem type results can be obtained. But when the roots are of unit modulus, it turns out that the asymptotic distributions are characterized in terms of the integrated integrals of Brownian motions. The key idea in obtaining these results lies in analyzing the order of magnitude of the observed Fisher's information matrix. Note that the LSE of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ can be expressed as

$$\boldsymbol{b}_n = (\sum_{t=1}^{n} \boldsymbol{y}_{t-1} \boldsymbol{y}_{t-1}{}^{\mathrm{T}})^{-1} \sum_{t=1}^{n} \boldsymbol{y}_{t-1} y_t, \tag{9}$$

where $\boldsymbol{y}_t = (y_t, \ldots, y_{t-p+1})^{\mathrm{T}}$ and $\boldsymbol{y}_0 = (0, \ldots, 0)^{\mathrm{T}}$. Similar to the estimation problem, different characteristic roots carry different information. By transforming the original nonstationary AR model into components according to their characteristic roots, Chan and Wei (1988) was able to derive the precise form of the limiting distributions. During the course of this investigation, they also obtained an important result about the weak convergence of stochastic integrals (Theorem 2.4 in Chan and Wei (1988)) which is of independent interest and has many applications in different areas (Kurtz and Protter (1991)). In addition, Chan and Wei also showed that different com-

ponents are asymptotically uncorrelated and, as a result, a joint limiting law can be established. Specifically, using the notation of Chan and Wei (1988), the following theorem was established.

Theorem 3 *Assume that* $\{y_t\}$ *follows (7) with the characteristic polynomial satisfying (8) and the innovation sequence* $\{\epsilon_t\}$ *satisfying (3). Then as* $n \to \infty$,

$$Q^{\mathrm{T}} G_n^{\mathrm{T}}(\boldsymbol{b}_n - \boldsymbol{\beta}) \to_D ((F^{-1}\boldsymbol{\xi})^{\mathrm{T}}, (\widetilde{F}^{-1}\boldsymbol{\eta})^{\mathrm{T}}, (H_1^{-1}\boldsymbol{\zeta}_1)^{\mathrm{T}}, \ldots, (H_\ell^{-1}\boldsymbol{\zeta}_\ell)^{\mathrm{T}}, N^{\mathrm{T}})^{\mathrm{T}},$$

where $(F, \boldsymbol{\xi})$, $(\widetilde{F}, \boldsymbol{\eta})$, $(H_1, \boldsymbol{\zeta}_1), \ldots, (H_\ell, \boldsymbol{\zeta}_\ell)$, N, G_n *and* Q *are independent and defined in equations (3.2) and (3.3) and Theorem 2.2 of Chan and Wei (1988).*

This result of Chan and Wei paved the way to the analysis of nonstationary processes and since then numerous extensions have been conducted. Jeganathan (1991) generalized this idea to the near-integrated situations where the limiting distributions of the LSE are expressed in terms of iterated integrals of Ornstein–Uhlenbeck processes. For the case when the underlying model has long memory, Chan and Terrin (1995) extended this result to functionals of fractional Brownian motions, while Ling and Li (1998) considered the case when the innovations are modeled by GARCH processes. Extensions of this result to vector AR processes are given in Tsay and Tiao (1990) and to processes with deterministic trends in Chan (1989). On the econometric front, Theorem 2.4 in Chan and Wei (1988) provides a fundamental tool in analyzing cointegrated systems. Comprehensive reviews on cointegrated vector autoregressions are given in Johansen (2008).

Beyond limiting distributions, another interesting issue is residual analysis for unit root series. Lee and Wei (1999) considered the stochastic regression model (1):

$$y_{nt} = \boldsymbol{\beta}_n^{\mathrm{T}} \boldsymbol{x}_{nt} + r_{nt} + \epsilon_{nt}, \quad 1 \le t \le n, \tag{10}$$

where $\boldsymbol{\beta}_n$ are unknown parameters, $\boldsymbol{x}_{nt}$ are observable random vectors and r_{nt} are random variables which they called "model bias." This model can be considered an extension of (1) as it encompasses both stochastic regressions and autoregressive time series. Let $\boldsymbol{b}_n$ denote the LSE of $\boldsymbol{\beta}_n$ by regressing y on $\boldsymbol{x}$ ignoring r, and let the residual be defined as $\widetilde{\epsilon}_{nt} = y_{nt} - \boldsymbol{b}_n^{\mathrm{T}} x_{nt}$. Consider the residual empirical process

$$\widehat{Y}_n(u) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} [I(H_n(\widetilde{\epsilon}_{nt}) \le u) - u], \tag{11}$$

where H_n is the underlying distribution of $\{\epsilon_{nt}\}$. Under certain regularity conditions on H_n and the growth rates of the orders of the model, Lee and Wei (1999) showed that for a Gaussian stationary AR(∞) model, under the setting of a null hypothesis $K_0 : H(\cdot) = \Phi(\cdot)$ and a contiguous sequence of alternatives $K_n : H_n(\cdot) = (1 - \gamma/\sqrt{n})\Phi(\cdot) + (\gamma/\sqrt{n})H(\cdot)$, where H is a

distribution function with mean zero and variance 1, the following theorem holds.

Theorem 4 *Under K_n, the residual empirical process $\widehat{Y}_n(u)$ defined in (11) converges weakly to a Gaussian process Y with mean and covariance*

$$\begin{aligned} \mathrm{E}Y(u) &= -\gamma(u - H \circ \Phi^{-1}(u)), \\ \mathrm{Cov}(Y(u), Y(v)) &= u \wedge v - uv - 0.5\phi(\Phi^{-1}(u))\Phi^{-1}(u)\phi(\Phi^{-1}(v))\Phi^{-1}(v), \end{aligned}$$

where $0 \le u, v \le 1$. Here Φ and ϕ denote the cumulative distribution function and the density function of a standard normal random variable, respectively.

In particular, for nonstationary AR(p) models, following the notation used in Lee and Wei (1999), let (W_1, W_2) be a mean zero two-dimensional Gaussian process with covariance structure such that for all $s, t \in [0, 1]$

$$\begin{aligned} \mathrm{Cov}(W_1(s), W_1(t)) &= s \wedge t - st, \\ \mathrm{Cov}(W_2(s), W_2(t)) &= s \wedge t, \\ \mathrm{Cov}(W_1(s), W_2(t)) &= (t/\sigma) \int_{-\infty}^{G^{-1}(s)} x \, dG(x), \end{aligned} \tag{12}$$

where G is the distribution function of the innovation sequence $\{\epsilon_t\}$ in (7). The following weak convergence result of the residual empirical processes was established in Lee and Wei (1999) (see also Ling (1998)).

Theorem 5 *Consider a nonstationary AR(p) model satisfying (7) with a characteristic root of 1 with multiplicity $a \ge 1$, and an iid innovation sequence $\{\epsilon_t\}$ with mean zero, variance $0 < \sigma^2 < \infty$ and continuous distribution function G. Then as $n \to \infty$,*

$$\widehat{Y}_n(u) \to_D \; W_1(u) + \sigma(F^{-1}\boldsymbol{\xi})^{\mathrm{T}} \boldsymbol{\eta} G^{\mathrm{T}}(G^{-1}(u)),$$

where (W_1, W_2) is the two-dimensional Gaussian process defined in (12); $F_0 = \sigma W_2$, $F_1 = \int_0^1 F_0(s)\, ds$, $F_j = \int_0^1 F_{j-1}(s)\, ds$, $j = 2, \ldots, a$; $\boldsymbol{\xi} = (\int_0^1 F_{a-1}(s)\, dW_2(s), \ldots, \int_0^1 F_0(s)\, dW_2(s))^{\mathrm{T}}$; $\boldsymbol{\eta} = (F_a(1), \ldots, F_1(1))^{\mathrm{T}}$; and F is the matrix whose (j, l)th entry is $\sigma_{jl} = \int_0^1 F_{j-1}(s) F_{l-1}(s)\, ds$.

This theorem indicates that the residual empirical process for an unstable AR(p) model with roots of 1 does not converge to a Brownian bridge as in the stable case. As a result, Lee and Wei recommended under such a situation one should conduct a unit root test before using conventional methods such as the Kolmogorov–Smirnov test. Recently, Chan and Ling (2008) considered the residual problem when the innovation sequence $\{\epsilon_t\}$ possesses a long-memory structure. They showed that a result similar to Theorem 2 can be derived, but the driving process becomes a fractional Brownian motion instead of a standard Brownian motion; further details can be found in Chan and Ling (2008).

2.3 Model selection

Consider the stochastic regression model (1) again. A natural problem is to study the performance of the LSEs in prediction and formulate the so-called predictive principle for model selection. By analyzing the order of the cumulative predictive errors

$$C_n = \sum_{k=1}^{n} (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_k - \boldsymbol{b}_{k-1}^{\mathrm{T}} \boldsymbol{x}_k)^2 = \sum_{k=1}^{n} (\widehat{\epsilon}_k - \epsilon_k)^2,$$

where $\widehat{\epsilon}_k = y_k - \widehat{y}_k = y_k - \boldsymbol{b}_{k-1}^{\mathrm{T}} \boldsymbol{x}_k$ is the one-step prediction error, Wei (1987) observed that the term C_n plays a crucial role for order selection.

Theorem 6 *Consider the regression model (1) with $\{\epsilon_i\}$ satisfying assumption (3). Assume that*

$$\boldsymbol{x}_n^{\mathrm{T}} \left(\sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^{\mathrm{T}} \right)^{-1} \boldsymbol{x}_n \to v \ \text{a.s. as } n \to \infty,$$

where v is a nonnegative random variable. Then

$$(1 - v) C_n + \sum_{k=1}^{n} [(\boldsymbol{b}_n - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{x}_k]^2 \sim n v \sigma^2 \ \text{a.s.}$$

on the set $\{1 > v > 0,\ C_n \to \infty\}$ and

$$C_n + \sum_{k=1}^{n} [(\boldsymbol{b}_n - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{x}_k]^2 \sim \sigma^2 \log \det \left(\sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^{\mathrm{T}} \right) \ \text{a.s.}$$

on the set $\{v = 0,\ C_n \to \infty,\ \lambda_{\min}(n) \to \infty\}$, where $\lambda_{\min}(n)$ denotes the minimum eigenvalue of the design matrix $\sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^{\mathrm{T}}$.

The proof of this result relies on the local martingale convergence theorem and follows the argument of Lai and Wei (1982a). With this result, a new order selection criterion was proposed in Wei (1987) for a nonstationary AR(p) model.

Theorem 7 *Assume that the autoregressive model (7) has roots equal to or bigger than 1 in magnitude (i.e., $b = d_1 = \cdots = d_\ell = 0$) and assume that $\beta_p \neq 0$ for an unknown p, but $r \geq p$ is given. Let $\boldsymbol{y}_n = (y_1, \ldots, y_{n-r+1})^{\mathrm{T}}$. Then*

$$\widehat{a}_n = [\log \det (\sum_{k=p}^{n} \boldsymbol{y}_k \boldsymbol{y}_k^{\mathrm{T}}) / \log n - r]^{1/2} \to a \ \textit{in probability.}$$

By means of the estimator $\widehat{a}_n$, one can determine how many times to difference an integrated time series to achieve stationarity when the exact order p

is unknown, but an upper bound r of the order is given. After differencing the integrated series $\widehat{a}_n$ times, one can then apply the traditional Akaike information criterion (AIC) or the Bayesian information criterion (BIC) for order selection. In other words, this theorem can be used to construct a two-step order selection procedure.

With this result, one can further pursue the notion of predictive least squares (PLS) in model selection. Wei (1992) reconsidered (1) and examined the conventional model selection criterion

$$\log \widehat{\sigma}_n^2 + c_n/n, \tag{13}$$

where n is the sample size and $\widehat{\sigma}_n^2$ is the residual variance after fitting the model based on $\boldsymbol{x}$, and c_n is a nonnegative random variable that measures the complexity of the model chosen, which is proportional to the number of parameters. Common criteria such as the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) fall within this setting. Motivated by (13), the idea of the predictive least squares criterion (PLS) criterion,

$$\mathrm{PLS}(\boldsymbol{x}) = \sum_{i=m+1}^{n} (y_i - \boldsymbol{b}_{i-1}^{\mathrm{T}} \boldsymbol{x}_i)^2, \tag{14}$$

was introduced and the notion of the so-called Fisher's information criterion (FIC) was introduced in Wei (1992), who showed that the FIC is equivalent to

$$\mathrm{FIC}(M) = n\widehat{\sigma}_n^2 + \widetilde{\sigma}_n^2 \log\det\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right), \tag{15}$$

where M is the model with design vector $\boldsymbol{x}_i$, and $\widehat{\sigma}_n^2$ and $\widetilde{\sigma}_n^2$ are variance estimators based on the model M and the full model, respectively. For a linear regression model with Gaussian errors, the conditional Fisher's information matrix is simply $\sigma^{-2} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$, which can be interpreted as the amount of information about the underlying unknown parameter. The FIC expression in (15) replaces the second quantity in the conventional criterion (13), which is proportional to the topological dimension of the model selected as reflected by c_n, by the second quantity of (15), which is proportional to the logarithm of the statistical information that is contained in M as reflected by the conditional Fisher's information matrix. This insight enables one to further link up PLS with FIC via

$$\mathrm{PLS} \sim n\widehat{\sigma}_n^2 + \sigma^2 \log\det\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right).$$

Replacing σ^2 on the right-hand side by an estimator, PLS is simply FIC. In summary, the following result was established in Wei (1992).

Theorem 8 *Assume either the stochastic regression model (1) holds with certain regularity conditions being satisfied, or that the nonstationary time series model (7) holds with a characteristic polynomial satisfying (8) together with certain regularity conditions (see Theorem 5.1.1 of Wei (1992)) being satisfied. Then the FIC is strongly consistent.*

This theorem has many important consequences. It is related to the order selection criterion studied in Pötscher (1989) and the predictive minimum description length idea used in Rissanen (1986). A similar idea was proposed by Phillips (1995) in a Bayesian setting, where it is known as the posterior information criterion (PIC). A thorough discussion on this subject can be found in the 1995 themed issue of volume 69 of the *Journal of Econometrics* entitled "Bayesian and Classical Econometric Modeling of Time Series." Further discussions about Bayesian unit root inference were given in Kadane et al. (1996). Ing and Wei (2003, 2005) built on Wei's idea to study the so-called same realization prediction principle for order selection of AR time series. They showed that the AIC is asymptotically efficient for same realization prediction problems.

3 Miscellaneous Developments and Conclusion

Although the field of time series in general and the subject of unit root time series in particular have been experiencing rigorous developments for the last few decades, there are still many open and interesting areas to be explored. In nonstationary time series, the recent popularity of long-memory models remains an open field; see Chan and Terrin (1995) and Buchmann and Chan (2007). Equally important is the area of empirical likelihood inference for time series. In addition to the maximum likelihood estimate, nonparametric procedures like empirical likelihood are gaining popularity and extension to this area is likely to be important; for related literature see Chuang and Chan (2002) and Chan and Ling (2006). Another area of importance is inference for infinite variance models. Here, many of the LSE-type results are no longer valid and an entirely new asymptotic theory needs to be established; see, for example, Chan and Tran (1989), Knight (1989, 1991), Phillips (1990), Daviset al. (1992) and Chan et al. (2006).

In summary, this review only offers a modest account of some of the developments of unit root time series. As witnessed by the aforementioned discussions, the subject of unit root time series has been manifested into different forms and found applications in areas such as engineering and econometrics. Through asymptotic inference, the scope of unit root time series has been greatly broadened and many profound challenges still remain to be resolved, which in turn will push the subject into new frontiers.

References

Anderson, T.W. and Taylor, J. (1979): Strong consistency of least squares estimators in dynamic models. *Annals of Statistics* **7**, 484–489.

Arató, M. (1982): Linear stochastic systems with constant coefficients: a statistical approach. *Lecture Notes in Control and Information Sciences* **45**, Springer, Berlin.

Basawa, I.V., Mallik, A.K., McCormick, W.P., Reeves, J.H. and Taylor, R.L. (1991): Bootstrapping unstable first-order autoregressive processes. *Annals of Statistics* **19**, 1098–1101.

Bobkoski, M.J. (1983): *Hypothesis testing in nonstationary time series*. Ph.D. thesis, Department of Statistics, University of Wisconsin, Madison.

Buchmann, B. and Chan, N.H. (2007): Inference for nearly unstable processes under strong dependence. *Annals of Statistics* **35**, 2001–2017.

Chan, N.H. (1988): The parameter inference for nearly nonstationary time series. *Journal of the American Statistical Association* **83**, 857–862.

Chan, N.H. (1989): On the nearly nonstationary seasonal time series. *Canadian Journal of Statistics* **17**, 279–284.

Chan, N.H. (1989): Asymptotic inference for unstable autoregressive time series with drifts. *Journal of Statistical Planning and Inference* **23**, 301–312.

Chan, N.H. (2002): *Time Series: Applications to Finance*. Wiley, New York.

Chan, N.H. and Ling, S.Q. (2006): Empirical likelihood for GARCH models. *Econometric Theory* **22**, 403–428.

Chan, N.H. and Ling, S.Q. (2008): Residual empirical processes for long-memory time series. *Annals of Statistics* to appear.

Chan, N.H. and Terrin, N.C. (1995): Inference for unstable long-memory processes with applications to fractional unit root autoregressions. *Annals of Statistics* **23**, 1662–1683.

Chan, N.H. and Tran, L.T. (1989): On the first-order autoregressive process with infinite variance. *Econometric Theory* **5**, 354–362.

Chan, N.H. and Tsay, R.S. (1996): Asymptotic inference for non-invertible moving average time series. *Journal of Time Series Analysis* **17**, 1–18.

Chan, N.H. and Wei, C.Z. (1987): Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics* **15**, 1050–1063.

Chan, N.H. and Wei, C.Z. (1988): Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16**, 367–401.

Chan, N.H., Peng, L. and Qi, Y. (2006): Quantile inference for nearly nonstationary autoregressive time series with infinite variance. *Statistica Sinica* **16**, 15–28.

Chuang, C.S. and Chan, N.H. (2002): Empirical likelihood for autoregressive models, with applications to unstable time series. *Statistica Sinica* **12**, 387–407.

Cox, D.D. and Llatas, I. (1991): Maximum likelihood type estimation for nearly nonstationary autoregressive time series. *Ann. Statist.* **19**, 1109–1128.

Cryer, J.D. and Ledolter, J. (1981): Small sample properties of the maximum likelihood estimator in the first-order moving average model. *Biometrika* **68**, 191–194.

Dahlhaus, R. (1985): *Data Tapers in Time Series Analysis*. Habilitation thesis, Universität-GHS, Essen.

Davis, R.A. and Dunsmuir, W.T.M. (1996): Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. *Econometric Theory* **12**, 1–29.

Davis, R.A., Knight, K. and Liu, J. (1992): M-estimation for autoregressions with infinite variance. *Stochastic Processes and their Applications* **40**, 145–180.

Dickey, D.A. and Fuller, W.A. (1979): Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**, 427–431.

Dickey, D.A., Bell, W.R. and Miller, R.B. (1986): Unit roots in time series models: tests and implications. *The American Statistician* **40**, 12–26.

Fuller, W.A. (1986): *Introduction to Statistical Time Series*, 2nd edition. Wiley, New York.

Ing, C.K. (2001): A note on mean-squared prediction errors of the least squares predictors in random walk models. *Journal of Time Series Analysis* **22**, 711–724.

Ing, C.K. and Wei, C.Z. (2003): On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis* **85**, 130–155.

Ing, C.K. and Wei, C.Z. (2005): Order selection for the same-realization prediction in autoregressive processes. *Annals of Statistics* **33**, 2423–2474.

Jeganathan, P. (1991): On the asymptotic behavior of least-squares estimators in AR time series with roots neat the unit circle. *Econometric Theory* **7**, 269–306.

Jeganathan, P. (1995): Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11**, 818–887.

Johansen, S. (2008): Cointegration: Overview and development *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 671–693. Springer, New York.

Kadane, J.B., Chan, N.H. and Wolfson, L. (1996): Priors for unit root models. *Journal of Econometrics* **75**, 99–111.

Knight, K. (1989): Limit theory for autoregressive-parameter estimates in an infinite-variance random walk. *Canadian Journal of Statistics* **17**, 261–278.

Knight, K. (1991): Limit theory for M-estimates in an integrated infinite variance process. *Econometric Theory* **7**, 186–199.

Kurtz, T.G. and Protter, P. (1991): Weak limit theorems for stochastic integrals and stochastic differential equations. *Annals of Probability* **19**, 1035–1070.

Lai, T.L., Robbins, H. and Wei, C.Z. (1978): Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Science of the USA* **75**, 3034–3036.

Lai, T.L. and Wei, C.Z. (1982a): Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics* **10**, 154–166.

Lai, T.L. and Wei, C.Z. (1982b): Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. *Journal of Multivariate Analysis* **13**, 1–23.

Lai, T.L. and Wei, C.Z. (1985): Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. *In: Krishnaiah, P.R. (Ed.): Multivariate Analysis VI*, 375–393. North-Holland, Amsterdam.

Larsson, R. (1998): Bartlett corrections for unit root test statistics. **19**, 425–438.

Lee, S.Y. and Wei, C.Z. (1999): On residual empirical processes of stochastic regression models with applications to time series. *Annals of Statistics* **27**, 237–261.

Ling, S. (1998): Weak convergence of the sequential empirical processes of residuals in nonstationary autoregressive models. *Annals of Statistics* **26**, 741–754.

Ling, S. and Li, W. K. (1998): Limiting distributions of maximum likelihood estimators for unstable autoregressive moving-average time series with general autoregressive heteroskedastic errors. *Annals of Statistics* **26**, 84–125.

Nielsen, B. (2005): Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms. *Econometric Theory* **21**, 534–561.

Perron, P. (1989): The great crash, the oil price shock and the unit root hypothesis. *Econometrica* **57**, 1361–1401.

Pham, D.T. (1992): Approximate distribution of parameter estimators for first-order autoregressive models. *Journal of Time Series Analysis* **13**, 147–170.

Phillips, P.C.B. (1987): Toward a unified asymptotic theory of autoregressions. *Biometrika* **74**, 535–547.

Phillips, P.C.B. (1990): Time series regressions with a unit root and infinite-variance errors. *Econometric Theory* **6**, 44–62.

Phillips, P.C.B. (1995): Bayesian prediction: a response. *Journal of Econometrics* **69**, 351–365.

Phillips, P.C.B. and Perron, P. (1988): Testing for a unit root in time series regression. *Biometrika* **75**, 335–346.

Pötscher, B.M. (1989): Model selection under nonstationary autoregressive models and stochastic linear regression models. *Annals of Statistics* **17**, 1257–1274.

Rissanen, J. (1986): Stochastic complexity and modeling. *Annals of Statistics* **14**, 1080–1100.

Saikkonen, P. and Luukkonen, R. (1993): Testing for a moving average unit root in autoregressive integrated moving average models. *Journal of the American Statistical Association* **88**, 596–601.

Stock, J.H. and Watson, M.W. (1988): Variable trends in economic time series. *Journal of Economic Perspective* **2**, 147–174.

Stock, J.H. (1994): Unit roots, structural breaks and trends. *In: Engle, R.F. and McFadden, D.L. (Eds.): Handbook of Econometrics, Vol IV*, 2739–2841. Elsevier, Amsterdam.

Tanaka, K. (1996): *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory* Wiley, New York.

Tiao, G.C. and Tsay, R.S. (1983): Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Annals of Statistics* **11**, 856–871.

Tsay, R.S. and Tiao, G.C. (1990): Asymptotic properties of multivariate nonstationary processes with applications to autoregressions. *Annals of Statistics* **18**, 220–250.

Watson, M.W. (1994): Vector autoregressions and cointegration. *In: Engle, R.F. and McFadden, D.L. (Eds.): Handbook of Econometrics, Vol IV*, 2843–2915. Elsevier, Amsterdam.

Wei, C.Z. (1985): Asymptotic properties of least squares estimates in stochastic regression models. *Annals of Statistics* **13**, 1498–1508.

Wei, C.Z. (1987): Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Annals of Statistics* **15**, 1667–1682.

Wei, C.Z. (1992): On predictive least squares principles. *Annals of Statistics* **20**, 1–42.

White, J.S. (1958): The limiting distribution of the serial correlation coefficient in the explosive case. *Annals of Mathematical Statistics* **29**, 1188–1197.

Fractional Cointegration

Willa W. Chen and Clifford M. Hurvich

Abstract We describe a variety of seimparametric models and estimators for fractional cointegration. All of the estimators we consider are based on the discrete Fourier transform of the data. This includes the ordinary least squares estimator as a special case. We make a distinction between Type I and Type II models, which differ from each other in terms of assumptions about initialization, and which lead to different functional limit laws for the partial sum processes. We compare the estimators in terms of rate of convergence. We briefly discuss the problems of testing for cointegration and determining the cointegrating rank. We also discuss relevant modeling issues, such as the local parametrization of the phase function.

1 Introduction

A collection of two or more time series, observed at equally-spaced time points, is *fractionally cointegrated* if there exists a non-null contemporaneous linear combination of the series, based on deterministic time-invariant weights, such that the linear combination is less persistent than any of the individual series, where persistence is measured in terms of the memory parameter, assumed here to be the same for all of the series. (In Section 3, we survey various generalizations to the case of unequal memory parameters of the observed series, which often arises in practice.) The most thoroughly-studied case to date is standard cointegration, in which the memory param-

Willa W. Chen
Department of Statistics, Texas A&M University, College Station, TX 77843, USA, e-mail: wchen@stat.tamu.edu

Clifford M. Hurvich
New York University, 44 W. 4'th Street, New York, NY 10012, USA, e-mail: churvich@stern.nyu.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_31,

eter is reduced from 1 to 0 by the linear combination. In general, there is no integer constraint, and no assumption of pre-knowledge, on the memory parameters of the original series or of the linear combination.

The seminal paper of Engle and Granger (1987) allowed for the fractional case, though fractional and standard cointegration have largely been developed separately in the literature. In this chapter, we will present a selective survey of the literature on representation, estimation and testing in fractional cointegration. As this literature is growing rapidly, we do not attempt to summarize it in its entirety, but rather we focus selectively on certain themes.

We will start by describing two different definitions of the order of integration of a time series, called the Type I and Type II definitions, which lead to different limit laws for partial sums. Both have been used in the fractional-cointegration literature. Next, we describe a variety of semiparametric models for fractional cointegration. Next, we consider estimation of the cointegrating parameters *i.e.*, the linear combinations that reduce the memory parameter. The various estimators can be classified according to the type of cointegration model assumed, the use of tapering and differencing, assumptions on the phase angle, the use of a fixed or increasing bandwidth, and the nature of the estimator: either a direct estimator of the cointegrating parameters alone, or a joint estimator of the cointegrating parameters together with the memory and other parameters. Finally, we discuss estimation of the cointegrating rank, and describe some results on testing for cointegration, focusing on residual-based tests and joint tests.

It should be emphasized that in this chapter, as in general, there is a link between the strength of assumptions and the strength of the theoretical results that can be established under these assumptions. Remembering this fact helps to put the existing results into an appropriate context, and may serve to prevent invidious comparisons in subsequent work.

2 Type I and Type II Definitions of $I(d)$

2.1 Univariate series

There are a variety of ways in which long memory can be defined. We start by considering univariate series, and then move on to consider vector series. Most definitions of long memory for a stationary series involve an asymptotic power law for some quantity, *e.g.*, the spectral density near zero frequency, the autocorrelation at large lags, the variance of partial sums with increasing aggregation, the moving average coefficients in an $MA(\infty)$ representation of the process. If the process $\{x_t\}$ is weakly stationary and invertible with spectral density f, autocovariance sequence $\{c_r\}_{r=-\infty}^{\infty}$, and moving

average representation $x_t = \sum_{k=0}^{\infty} a_k \epsilon_{t-k}$ where $\{\epsilon_t\}$ are white noise, the long-memory assumption may be written in terms of the memory parameter $d \in (-1/2, 1/2) - \{0\}$ in the following generally non-equivalent ways:

$$f(\omega) \sim K_1 \omega^{-2d} \text{ as } \omega \to 0^+,$$

$$c_r \sim K_2 r^{2d-1} \text{ as } r \to \infty,$$

$$\text{var} \sum_{t=1}^{n} x_t \sim K_3 n^{2d+1} \text{ as } n \to \infty,$$

$$a_k \sim K_4 k^{d-1} \text{ as } k \to \infty,$$

where $K_1 > 0$, K_2, $K_3 > 0$ and K_4 are constants. Connections between these properties are given in Robinson (1995b) and Taqqu (2003). All of these properties hold for fractional ARIMA models (Adenstedt 1974, Granger and Joyeux 1980, Hosking 1981), with $d \in (-1/2, 1/2) - \{0\}$, though semiparametric specifications of long memory typically assume one of these properties at the potential expense of the others. We will say that a weakly-stationary series has *short memory* if its spectral density satisfies the property given above with $d = 0$, *i.e.*, the spectral density tends to a positive constant as the frequency tends to zero.

We will say that a process is *integrated of order d*, denoted by $I(d)$, if it has memory parameter d. Thus, for a stationary and invertible series, $I(d)$ can be defined by any of the asymptotic power-law relations given above. Since in fractional cointegration the original series may be nonstationary, it is essential to be able to define $I(d)$ even for $d > 1/2$. It is also convenient to be able to define $I(d)$ in the non-invertible case, $d \leq -1/2$.

Hurvich and Ray (1995) worked with the spectral definition of long memory as given above, which remains valid in the non-invertible case. For all $d < 1/2$, they define a weakly stationary process to be $I(d)$ if its spectral density satisfies $f(\omega) \sim K_1 \omega^{-2d}$ as $\omega \to 0^+$, where $K_1 > 0$. In the non-stationary case $d > 1/2$, they define a process to be $I(d)$ if there exists a positive integer k such that the k'th ordinary difference of the series is $I(d-k)$, where $d - k \in (-1/2, 1/2)$. For example, if $\{x_t\}$ is a random walk, $x_t = x_{t-1} + \epsilon_t$ where $\{\epsilon_t\}$ is white noise, the ordinary first difference is $x_t - x_{t-1} = \epsilon_t$, which is $I(0)$, so $\{x_t\}$ is $I(1)$. In a semiparametric context, the definition (henceforth referred to as the Type I definition) has been used in papers on estimation of the memory parameter by Velasco (1999a, 1999b), and Hurvich and Chen (2000) and in papers on fractional cointegration by Chen and Hurvich (2003a, 2003b, 2006), and Velasco (2003).

Marinucci and Robinson (1999) present an alternative definition of $I(d)$ based on truncation, which we will call the Type II definition. Given the choice of a time origin, $t = 0$ (which plays an important role here), for any $d > -1/2$, the (Type II) $I(d)$ series is represented for $t \geq 1$ as

$$x_t = \sum_{k=0}^{t-1} \psi_k \eta_{t-k}, \tag{1}$$

where $\{\eta_t\}$ is a weakly stationary, zero-mean short-memory series, and for $d \neq 0$

$$\psi_k = \Gamma(k+d)/[\Gamma(k+1)\Gamma(d)] \sim Ck^{d-1} \text{ as } k \to \infty \ (C > 0),$$

while for $d = 0$, we define $\psi_0 = 1, \psi_k = 0$ $(k \neq 0)$. A more general formulation for ψ_k, with the same limiting behavior as above, can also be considered, subject to suitable regularity conditions. Even for nonzero $d \in (-1/2, 1/2)$ the series $\{x_t\}$ is not stationary, but it is asymptotically stationary in the sense that there exists a function g such that for $t > u$, $\text{cov}(x_t, x_u) \sim g(t-u)$ if $u/(t-u) \to 0$.

A representation equivalent to (1) is

$$x_t = (1-B)^{-d} \eta_t^* \quad (t \geq 1), \tag{2}$$

where B is the backshift operator, and $\eta_t^* = \eta_t$ for $t \geq 1$, $\eta_t^* = 0$ for $t \leq 0$. In (2), $(1-B)^{-d}$ is defined for all d through the formal expansion (see, e.g., Brockwell and Davis 1991, p. 522),

$$(1-B)^{-d} = \sum_{k=0}^{\infty} \psi_k B^k.$$

It follows that for $d > 1/2$, the Type-II $I(d)$ process, which is neither stationary nor asymptotically stationary in this case, can be written as the partial sum

$$x_t = u_1^* + u_2^* + \cdots + u_t^* \quad t \geq 1,$$

where

$$u_t^* = (1-B)^{1-d} \eta_t^* = \sum_{k=0}^{t-1} a_k \eta_{t-k},$$

with $a_k = \Gamma(k+d-1)/[\Gamma(k+1)\Gamma(d-1)]$, so that $\{u_t^*\}$ is Type-II $I(d-1)$. In this case $(d > 1/2)$, Marinucci and Robinson (2000), following earlier work of Akonom and Gourieroux (1987) and Silveira (1991), derived the weak limit of suitably normalized partial sums of $\{u_t^*\}$ assuming that $\{\eta_t\}$ has a linear representation with respect to an *iid* or absolutely regular sequence (see Pham and Tran 1985). The limiting process, which Marinucci and Robinson (2000) called Type II fractional Brownian motion, also known as Riemann-Liouville fractional Brownian motion, $W_H(\cdot)$ (with $H = d + 1/2$), is defined for any $H > 0$, whereas the so-called Type I fractional Brownian motion $B_H(\cdot)$ considered by Mandelbrot and Van Ness (1968) and others, is only defined for $0 < H < 1$, and even in this case the two processes are not equivalent.

Limit theory has also been developed for statistics based on $I(d)$ processes under the Type I definition. For example, Sowell (1990) obtained the limiting distribution of the OLS estimator in a regression of x_t on x_{t-1} for a process $\{x_t\}$ such that $(1-B)^d x_t = \epsilon_t$, where ϵ_t is iid with zero mean, with $d \in (0.5, 1.5)$. The limit distribution is a functional of Type-I fractional Brownian motion. (See the remark on Sowell's results in Marinucci and Robinson 1999).

Implications of the Type I and Type II definitions of $I(d)$ on properties of the discrete Fourier transform (DFT), semiparametric log-periodogram regression estimates of the memory parameter, and other statistics of interest were studied by Velasco (2007). There are noticeable differences in the properties of the DFTs even for $d < 1/2$, but these do not have any important impact on the properties of the memory parameter estimates unless $d \geq 1/2$. It is notable in this regard that, for the region $d \in [1/2, 3/4)$, assuming a Gaussian process, the (suitably normalized) log-periodogram regression estimate of d has been shown to be asymptotically normal in the Type-I case (Velasco 1999a), but no such result has been established as of yet in the Type-II case, owing perhaps to the much stronger asymptotic correlations of the normalized DFTs under Type-II compared to Type I $I(d)$ when $d \geq 1/2$.

2.2 Multivariate series

Here, we present a Type I definition of $I(d)$ for a q-dimensional vector process $\{x_t\}$. We focus on the stationary case, in which the memory parameters of the entries of the vector are all less than $1/2$. Extension to the nonstationary case is then done similarly as for the univariate Type I $I(d)$ case.

Recall that for any weakly stationary real-valued q-vector process $\{x_t\}$ with spectral density f, we have $f(-\omega) = \overline{f(\omega)}$ for all $\omega \in [-\pi, \pi]$, and the lag-h autocovariance matrix is

$$\widetilde{\Gamma}(h) = \mathbb{E}[x_{t+h} x_t'] = \int_{-\pi}^{\pi} f(\omega) e^{ih\omega} d\omega \ ,$$

so that $\widetilde{\Gamma}(h)\widetilde{\Gamma}'(-h)$, where the superscript prime denotes transposition.

A weakly stationary q-vector process $\{x_t\}$ is (Type I) $I(d_1, \ldots, d_q)$ if its spectral density matrix f satisfies

$$f(\omega) \sim \Lambda G \Lambda^*, \qquad \omega \to 0^+, \tag{3}$$

where G is a nonnegative definite real symmetric matrix with nonzero diagonal entries not depending on ω,

$$\Lambda = diag\left(e^{i\phi_1} |\omega|^{-d_1}, \ldots, e^{i\phi_q} |\omega|^{-d_q}\right)$$

with $d_k < 1/2$, for $k = 1, \ldots, q$, Λ^* is the conjugate transpose of Λ, $\phi_1 = 0$, and the phase angles ϕ_k $(k = 2, \ldots, q)$ are real. The standardization $\phi_1 = 0$ is made for the sake of identifiability, since f in (3) remains unchanged if Λ is multiplied by a complex number with modulus 1. The series $\{x_t\}$ may or may not be cointegrated, but if it is then G must be singular.

Several special cases of the model have been considered in the literature.

Case 1: The case where all the phase angles are zero ($\phi_1 = \phi_2 = \cdots = \phi_q = 0$) was assumed by Christensen and Nielsen (2006). One way this could happen would be if the spectral density $f(\omega)$ were real for all $\omega \in [-\pi, \pi]$. In this case, we would have $\widetilde{\Gamma}(h) = \widetilde{\Gamma}(-h)$, so that the lag-$h$ cross-covariance of two entries of $\{x_t\}$ would be the same regardless of which entry leads. This would clearly entail a loss of generality. In the case of an ARFIMA model, one would obtain $\phi_1 = \phi_2 = \cdots = \phi_q = 0$ if and only if $d_1 = d_2 = \cdots = d_q$.

Case 2: The case $\phi_k = (\pi/2)d_k$ was considered in Robinson and Yajima (2002), Shimotsu (2006) and Lobato (1999) (though for convenience of estimation, Lobato (1999) subsequently changed to the assumption that all phase angles are zero). The assumption $\phi_k = (\pi/2)d_k$ is satisfied by the fractional ARIMA model. In view of the identifiability problem discussed above, the assumption is equivalent to $\phi_k = (\pi/2)(d_k - d_1)$.

Case 3: A more general case is $\phi_k = (\pi/2 - \gamma_k)(d_k - d_1)$ where $\gamma_2, \ldots, \gamma_q$ are real numbers satisfying the constraints that the resulting values of ϕ_k are in $(-\pi, \pi)$. Note that these constraints depend on the memory parameters. This case was considered by Robinson (2006) in a bivariate setting.

Chen and Hurvich (2003a,b, 2006) assumed a model that is ultimately equivalent to (3), but they made an additional assumption, which was not needed for their theoretical results, that the phase functions in the transfer function for the linear representation of the fractionally differenced series are continuous (and therefore zero) at zero frequency. If this assumption is removed, then the model becomes equivalent to (3). Such a model is more general than those obtained in Cases 1, 2 and 3 above, since the ϕ_k may be free of, or depend nonlinearly on $(d_1, \ldots, d_q)$. Chen and Hurvich (2003a,b, 2006) obtained the limiting asymptotic normality of the discrete Fourier transform of a multivariate Type I $I(d_1, \ldots, d_q)$ series for a fixed set of Fourier frequencies, assuming that the series has a linear representation with respect to an *iid* sequence.

A multivariate Type II definition for $I\,(d_1, \ldots, d_q)$ was given by Robinson and Marinucci (2001, 2003), and Marinucci and Robinson (2000), also used by Marmol and Velasco (2004) and Nielsen and Shimotsu (2007), as $diag((1 - B)^{d_1}, \ldots, (1 - B)^{d_q})X_t = u_t I\{t \geq 1\}$ for $t = 1, 2, \ldots$ where $\{X_t\}$ is the observed series, $\{u_t\}$ is a stationary short-memory series with zero mean, and $d_k > -1/2$, $k = 1, \ldots, q$. Limit theory for partial sums of such processes was developed by Marinucci and Robinson (2000).

3 Models for Fractional Cointegration

As stated in the introduction, a $(q\times 1)$ $I(d_0,\ldots,d_0)$ series $\{x_t\}$ is *cointegrated* if there exists a vector $\alpha \neq 0$, such that $u_t = \alpha' x_t$ is $I(d_u)$ where $d_u < d_0$. Any such vector α is called a *cointegrating vector*. There may be up to $q-1$ linearly independent cointegrating vectors, and the corresponding contemporaneous linear combinations may have different memory parameters. The number of linearly independent cointegrating vectors, r, with $1 \leq r < q$, is called the *cointegrating rank*. It is assumed in the above definitions, as in the work of Engle and Granger (1987), and Chen and Hurvich (2003a, 2003b, 2006), among others, that all entries of the observed series have the same memory parameter. For a general $I(\widetilde{d}_1,\cdots,\widetilde{d}_q)$ series, the concept of cointegration requires a more careful definition, and several different ones have been provided, as described in Robinson and Yajima (2002). By any of these definitions, a necessary condition for cointegration is that at least two of $\widetilde{d}_1,\ldots,\widetilde{d}_q$ are equal. Robinson and Yajima (2002, Definition 2) start by partitioning the entries of the observed series into blocks with equal memory parameters within each block, and unequal memory parameters across blocks. (They also provide a data-driven procedure for constructing such a partition, which is successful at this task with probability approaching 1). The $(q \times 1)$ series is cointegrated if for some block the entries of that block are cointegrated in the sense defined earlier. The cointegrating rank of the entire $(q\times 1)$ series is the sum of the cointegrating ranks of the blocks. Robinson and Marinucci (2003) use a different definition, under which the $(q \times 1)$ series $\{x_t\}$ is cointegrated if there exists $\alpha \neq 0$ such that $\alpha' x_t = u_t$ is $I(d_u)$ with $d_u < \min(\widetilde{d}_1,\ldots,\widetilde{d}_q)$.

We will present some semiparametric Type I and Type II models for fractional cointegration. We start with the Bivariate Type I model $y_t = \beta x_t + u_t$, where the observed series $\{x_t\}$ and $\{y_t\}$ are both $I(d)$, the unobserved series $\{u_t\}$ is $I(d_u)$, and $d_u < d$, so that $\{(x_t, y_t)'\}$ is fractionally cointegrated. We do not require that $\{x_t\}$ and $\{u_t\}$ be mutually independent. The stationary, positive-memory-parameter case d, $d_u \in (0, 1/2)$ was considered by Robinson (1994). Chen and Hurvich (2003a) assumed that the original series $\{X_t\}$, $\{Y_t\}$ $\{U_t\}$ are potentially nonstationary, but after $p-1$ ordinary differences they yield the stationary (but potentially noninvertible) series $\{x_t\}$, $\{y_t\}$, $\{u_t\}$ satisfying $y_t = \beta x_t + u_t$, with d, $d_u \in (-p+1/2, 1/2)$, $p \geq 1$. For example, if $p = 1$, the range would be $(-1/2, 1/2)$, i.e., the stationary invertible case. If $p = 2$, the range would be $(-3/2, 1/2)$, which allows for noninvertibility induced by (potentially unintentional) overdifferencing.

Chen and Hurvich (2006) proposed the Type I fractional common components model for the $(q \times 1)$ observed series $\{y_t\}$ with cointegrating rank r $(1 \leq r < q)$, and s cointegrating subspaces $(1 \leq s \leq r)$, given by

$$y_t = \mathbf{A}_0 u_t^{(0)} + \mathbf{A}_1 u_t^{(1)} + \cdots + \mathbf{A}_s u_t^{(s)}, \tag{4}$$

where $\mathbf{A}_k$ $(0 \leq k \leq s)$ are unknown $q \times a_k$ full-rank matrices with $a_0 = q - r$ and $a_1 + \cdots + a_s = r$ such that all columns of $\mathbf{A}_0, \ldots, \mathbf{A}_s$ are linearly independent, $\{u_t^{(k)}\}$ $k = 0, \ldots, s$, are unobserved $(a_k \times 1)$ processes with memory parameters $\{d_k\}_{k=0}^s$ with $-p + 1/2 < d_s < \cdots < d_0 < 1/2$. Thus, each entry of $\{y_t\}$ is $I(d_0)$, and the space $\mathbb{R}^q$ can be decomposed into a direct sum of orthogonal *cointegrating subspaces* of dimension a_k $(k = 0, \ldots, s)$ such that any vector α in the k'th cointegrating subspace yields a linear combination $\alpha' y_t$ which is $I(d_k)$ with $d_k < d_0$ if $k > 0$. The series $\{y_t\}$ is stationary, the result of $p - 1$'th differencing of an original, potentially nonstationary series $\{Y_t\}$, with $p \geq 1$. The case of standard cointegration could be obtained, for example, by taking $p = 2$, $s = 1$, $d_0 = 0$, $d_1 = -1$, and $1 \leq r \leq q - 1$. Equation (4) can be written as

$$y_t = \mathbf{A} z_t, \tag{5}$$

where $z_t = vec\left(u_t^{(0)}, \ldots, u_t^{(s)}\right)$ and $\mathbf{A} = \left[\mathbf{A}_0 \cdots \mathbf{A}_s\right]$. Chen and Hurvich (2006) make additional assumptions on the spectral density of $\{z_t\}$. These assumptions guarantee that $\{z_t\}$ is not cointegrated, and that the spectral density of $\{y_t\}$ satisfies (3) with G singular, which is in turn a necessary and sufficient condition for cointegration in a stationary process such that all components have the same memory parameter. The methodology presented in Chen and Hurvich (2006) does not require either r or s to be known.

Robinson and Marinucci (2003) considered a Type-II model with cointegrating rank 1 for a non-differenced $q \times 1$ series $\{Z_t\}$ which is partitioned as $\{Z_t\} = \{(Y_t, X_t')'\}$ with $\{Y_t\}$ 1×1 and $\{X_t\}$ $(q-1) \times 1$ such that $Y_t = \beta' X_t + U_t$, where β is an unknown $(q-1) \times 1$ cointegration parameter, and $\{U_t\}$ has a smaller memory parameter than the minimum memory parameter of the entries of $\{Z_t\}$. Thus, $(1, -\beta')'$ is a cointegrating vector for $\{Z_t\}$. The need to specify one of the entries of $\{Z_t\}$ as the response variable may cause difficulties when $q \geq 3$ since there is no guarantee in general that all entries of $\{Z_t\}$ appear in a cointegrating relationship with at least one of the other entries. Thus if a randomly chosen component of $\{Z_t\}$ is labeled as the response variable, there is no guarantee that the regression model above will hold, and in any case regression-type estimators of the parameter will not be invariant to this choice.

3.1 Parametric models

The models for cointegration considered above are all semiparametric, in that the spectral density is only specified in a neighborhood of zero frequency. Parametric models are also of interest, in which the time-series dynamics of the series is fully determined by a finite set of fixed, unknown parameters, although the distribution of the innovations may not be parametrically

specified. For reasons of brevity, we will not present a detailed discussion of parametric models for cointegration, or of the properties of estimators in such models, but we provide some references here. Gaussian Maximum likelihood estimation of a cointegrated fractional ARIMA model was considered by Dueker and Startz (1998). Maximum likelihood estimation of the cointegrating parameter in a multivariate Type-I model with a known but not necessarily Gaussian distribution was considered by Jeganathan (1999). Estimation of the cointegrating parameter in a bivariate nonstationary Type-II model was considered using generalized least-squares by Robinson and Hualde (2003). Here, the rate of convergence for the estimator of the cointegrating parameter is n^{δ}, where the degree of cointegration (i.e. reduction in the memory parameter) δ is assumed to be greater than $1/2$. Properties of ordinary least-squares estimators were considered for Type-I autoregressive models with fractional errors by Chan and Terrin (1995). A parametric cointegration model of Granger (1986) was followed up in a fractional context by Breitung and Hassler (2002), and applied by Davidson (2002) and Dolado, Gonzalo and Mayoral (2003).

4 Tapering

Tapering is the multiplication of an observed series by a sequence of constants (the *taper*) prior to Fourier transformation, in order to reduce bias in the periodogram. A cosine bell taper was used in Hurvich and Ray (1995), and was found to be especially helpful in both the noninvertible case ($d < -1/2$) and the nonstationary case ($d > 1/2$). A class of tapers due to Zhurbenko was used by Velasco (1999a, 1999b) for estimation of d in the nonstationary case. Hurvich and Chen (2000) chose to difference the data $p-1$ times to remove deterministic polynomial trends and to render the series stationary, and then proposed applying the p'th order taper $\{h_t^{p-1}\}$, where $h_t = (1/2)[1 - \exp\{i2\pi(t-1/2)/n\}]$, to handle the potential noninvertibility of the differenced series. They showed that this allows for Gaussian semiparametric estimation of d with a smaller variance inflation than incurred by the methodology of Velasco (1999b). Variance inflation in estimation of d can be removed entirely using the class of tapers of Chen (2006), and the resulting estimator is therefore comparable to the non-tapered exact Local Whittle estimator of Shimotsu and Phillips (2005). In a cointegration context, tapering was used by Chen and Hurvich (2003a, 2003b, 2006), who used differencing together with the tapers of Hurvich and Chen (2000), and Velasco (2003). Since the taper of Hurvich and Chen (2000) is complex-valued, care must be taken in using fast Fourier transform software, since $h_t^{p-1}\exp(-i\lambda)$ is not the conjugate of $h_t^{p-1}\exp(i\lambda)$.

5 Semiparametric Estimation of the Cointegrating Vectors

We first consider direct estimators of the cointegrating parameter, i.e., estimators which can be constructed without estimating other nuisance parameters, such as memory parameters and phase parameters. The limiting distributions of these estimators have been shown to be non-Gaussian in most cases, which is somewhat undesirable from the point of view of practical usefulness. We focus here on rates of convergence. We start with the bivariate model

$$y_t = \beta x_t + u_t,$$

in the Type-I stationary positive-memory case $d, d_u \in (0, 1/2)$, $d_u < d$, as considered by Robinson (1994). Suppose we have n observations on the response and explanatory variables, $\{y_t\}_{t=1}^n$, $\{x_t\}_{t=1}^n$. Consider the ordinary least squares (OLS) estimator $\widehat{\beta}^{OLS}$ of β obtained from linear regression of $\{y_t\}$ on $\{x_t\}$. If it is known that $\{u_t\}$ has zero mean, the regression can be performed without an intercept. Robinson (1994) pointed out that if $\widehat{\beta}^{OLS}$ converges in probability, the probability limit must be

$$\beta + \mathrm{cov}(x_t, u_t)/\mathrm{var}(x_t).$$

Thus, as long as there is a nonzero contemporaneous correlation between x_t and u_t, $\widehat{\beta}^{OLS}$ will be an inconsistent estimator of β.

Robinson (1994) proposed a modification of $\widehat{\beta}^{OLS}$ that remains consistent even if $\mathrm{cov}(x_t, u_t) \neq 0$. To motivate the modification, it is helpful to consider the source of the inconsistency of $\widehat{\beta}^{OLS}$ from a frequency-domain perspective. Note that

$$\mathrm{cov}(x_t, u_t)/\mathrm{var}(x_t) = \int_{-\pi}^{\pi} f_{xu}(\omega)d\omega \Big/ \int_{-\pi}^{\pi} f_x(\omega)d\omega,$$

where f_x is the spectral density of $\{x_t\}$ and f_{xu} is the cross-spectral density of $\{(x_t, u_t)'\}$. By the Cauchy-Schwartz inequality, $f_{xu}(\omega) = O(\omega^{-(d+d_u)})$ as $\omega \to 0^+$, so that $f_{xu}(\omega)/f_x(\omega) \to 0$ as $\omega \to 0^+$. Thus, if ω lies in a neighborhood which shrinks to zero as the sample size n increases, the contribution to $\mathrm{cov}(x_t, u_t)$ from frequency ω is negligible compared to the contribution to $\mathrm{var}(x_t)$ from frequency ω. This motivates the narrowband least-squares estimator of β, given by Robinson (1994) as

$$\widehat{\beta}^{NBLS} = \sum_{j=1}^{m_n} Re\{J_{x,j}\bar{J}_{y,j}\} \Big/ \sum_{j=1}^{m_n} |J_{x,j}|^2,$$

a frequency-domain regression of y on x, where m_n is a bandwidth parameter, $m_n \to \infty$, $m_n/n \to 0$ as $n \to \infty$,

$$J_{z,j} = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} z_t \exp(i\omega_j t)$$

for any series $\{z_t\}$, and $\omega_j = 2\pi j/n$. It then follows from Robinson (1994) that $\widehat{\beta}^{NBLS} \xrightarrow{p} \beta$ as $n \to \infty$, under an additional assumption on the errors in the $MA(\infty)$ representation of $\{x_t\}$, and assuming that $d, d_u \in (0, 1/2)$. If, in violation of the assumption $m_n/n \to 0$, we take $m_n = \lfloor n/2 \rfloor$ (so that $\widehat{\beta}^{NBLS}$ is no longer in fact narrowband), then $\widehat{\beta}^{NBLS} = \widehat{\beta}^{OLS}$, where $\widehat{\beta}^{OLS}$ is computed from a regression that includes an intercept. If for this same choice of m_n the lower index of the summations in $\widehat{\beta}^{NBLS}$ were changed from 1 to 0, then we would again have $\widehat{\beta}^{NBLS} = \widehat{\beta}^{OLS}$, but in this case the OLS regression would not contain an intercept.

Robinson and Marinucci (2001) derived asymptotic properties of $\widehat{\beta}^{NBLS}$ (with $m_n \to \infty$, $m_n/n \to 0$) and $\widehat{\beta}^{OLS}$ for the bivariate model $Y_t = \beta X_t + U_t$, assuming that the observed series $\{X_t\}$ and $\{Y_t\}$ are nonstationary Type-II $I(d_X)$ with $d_X \geq 1/2$, and the unobserved errors $\{U_t\}$ are Type-II $I(d_U)$ with $d_U < d_X$. They show that $\widehat{\beta}^{OLS}$ and $\widehat{\beta}^{NBLS}$ based on the observed non-differenced data $\{X_t\}_{t=1}^n$, $\{Y_t\}_{t=1}^n$ are consistent, with a rate of convergence that depends on d_X and d_U, but not simply on $d_X - d_U$.

Chen and Hurvich (2003a) considered a Type-I bivariate model, and proposed $\widehat{\beta}^{CH}$, a modified version of $\widehat{\beta}^{NBLS}$ computed from tapered, differenced data. The original series $\{X_t\}$, $\{Y_t\}$ and $\{U_t\}$ are differenced $p-1$ times, resulting in stationary series $\{x_t\}$, $\{y_t\}$ with memory parameter d, and $\{u_t\}$ with memory parameter d_u, $d_u < d$, and $d, d_u \in (-p + 1/2, 1/2)$. The differencing transforms the original model $Y_t = \beta X_t + U_t$ into $y_t = \beta x_t + u_t$. The estimator is then given by

$$\widehat{\beta}^{CH} = \sum_{j=1}^{m_0} Re\{J_{x,j}^T \bar{J}_{y,j}^T\} / \sum_{j=1}^{m_0} |J_{x,j}^T|^2,$$

where $m_0 \geq 1$ is a fixed bandwidth parameter,

$$J_{z,j}^T = \sum_{t=1}^{n} h_t^{p-1} z_t \exp(-i\omega_j t),$$

and

$$h_t = (1/2)[1 - \exp\{i2\pi(t - 1/2)/n\}]. \tag{6}$$

Chen and Hurvich (2003a) showed that subject to suitable regularity conditions the rate of convergence of $\widehat{\beta}^{CH}$ is always n^{d-d_u} (the same as $n^{d_X - d_U}$), in the sense that $n^{d-d_u}(\widehat{\beta}^{CH} - \beta)$ converges in distribution. This uniformity of the rate of convergence of $\widehat{\beta}^{CH}$ (i.e. the dependence on $d - d_u$ alone), as contrasted with the nonuniformity of the rate of $\widehat{\beta}^{OLS}$ or $\widehat{\beta}^{NBLS}$, is due to the combination of tapering and the use of a fixed bandwidth m_0 in $\widehat{\beta}^{CH}$.

Though it may at first seem surprising that $\widehat{\beta}^{CH}$ can be consistent for β even though m_0 is fixed (after all, the use of a fixed number of frequencies in semiparametric estimation of the memory parameter would result in an inconsistent estimator), the consistency of $\widehat{\beta}^{CH}$ follows since it can be shown that each term of

$$\widehat{\beta}^{CH} - \beta = \sum_{j=1}^{m_0} Re\{J_{x,j}^T \bar{J}_{u,j}^T\} / \sum_{j=1}^{m_0} |J_{x,j}^T|^2$$

is $O_p(n^{d_u - d})$, and since there are a fixed number of terms.

We present here some comparisons of the rate of convergence of $\widehat{\beta}^{CH}$ with those of $\widehat{\beta}^{OLS}$ and $\widehat{\beta}^{NBLS}$, keeping in mind that the theory for the former estimate is based on a Type-I model, while that for the latter estimates is based on a Type-II model. Let d_X and d_U represent the memory parameters of the nondifferenced observed and error series, $\{X_t\}$ and $\{U_t\}$. All of our comparisons below are based on the assumption $d_X \geq 1/2$ (the observed series are nonstationary), as this is assumed in the results we will use from Robinson and Marinucci (2001). The rates of convergence for $\widehat{\beta}^{OLS}$ and $\widehat{\beta}^{NBLS}$ remain the same whether or not the zero frequency is excluded from the sums in their definitions. Since $d_X \geq 1/2$, $\widehat{\beta}^{OLS}$ is consistent. This follows from the Cauchy-Schwartz inequality and the fact that in this case

$$\sum_{t=1}^{n} U_t^2 / \sum_{t=1}^{n} X_t^2 \xrightarrow{p} 0.$$

For a given value of the difference of the memory parameters (i.e., the strength of the cointegrating relationship), $\widehat{\beta}^{CH}$, which is based on the tapered differences of order $p-1$ (and is invariant to additive polynomial trends of order $p-1$ in the levels) will be $n^{d_X - d_U}$-consistent. The comparison is separated into several cases, due to the non-uniform rates of convergence for $\widehat{\beta}^{OLS}$ and $\widehat{\beta}^{NBLS}$. All three estimators converge at the same rate when $d_U > 0$, $d_U + d_X > 1$, and also when $d_U = 0$, $d_X = 1$. The estimators $\widehat{\beta}^{CH}$ and $\widehat{\beta}^{OLS}$ converge at the same rate when $d_U = 0$, $d_X > 1$, though Robinson and Marinucci (2001) do not present a rate of convergence for $\widehat{\beta}^{NBLS}$ in this case. In the remaining cases, $\widehat{\beta}^{CH}$ has a faster rate of convergence than the other two estimators. We present the comparisons in terms of the improvement factor, given as the ratio of the rates of convergence. For example, if another estimator is n^{γ}-consistent with $\gamma < d_X - d_U$ then the improvement factor for $\widehat{\beta}^{CH}$ relative to the other estimator is $n^{d_X - d_U - \gamma}$. For the case $d_U > 0$, $d_U + d_X = 1$, the improvement factor for $\widehat{\beta}^{CH}$ relative to $\widehat{\beta}^{OLS}$ is $\log n$, and the improvement factor for $\widehat{\beta}^{CH}$ relative to $\widehat{\beta}^{NBLS}$ is $\log m_n$. For the case $d_U \geq 0$, $d_U + d_X < 1$, the improvement factor for $\widehat{\beta}^{CH}$ relative to $\widehat{\beta}^{OLS}$ is $n^{1-d_U-d_X}$, and the improvement factor for $\widehat{\beta}^{CH}$ relative to $\widehat{\beta}^{NBLS}$ is $m_n^{1-d_U-d_X}$. In the latter two cases, the slower the rate of increase of m_n,

the less inferior is the performance of $\widehat{\beta}^{NBLS}$ compared to that of $\widehat{\beta}^{CH}$. This helps to justify the use of a fixed bandwidth m_0 in $\widehat{\beta}^{CH}$. It can be seen that the spectral density and periodogram matrices at the *very* low frequencies (e.g., $\omega_1, \ldots, \omega_{m_0}$ with m_0 fixed) play a key role in fractional cointegration.

For a multivariate series, Robinson and Marinucci (2003) considered properties of NBLS and OLS estimators of the cointegrating parameter β in the multiple regression model

$$Y_t = \beta' X_t + U_t$$

described above. The estimators are

$$\widehat{\beta}^{NBLS} = \widehat{F}_{XX}(1, m_n)^{-1} \widehat{F}_{XY}(1, m_n)$$

and $\widehat{\beta}^{OLS}$ given by a similar formula with $(1, m_n)$ replaced by $(1, \lfloor n/2 \rfloor)$ or $(0, \lfloor n/2 \rfloor)$, where

$$\widehat{F}_{ab}(\ell, k) = \sum_{j=\ell}^{k} Re(J_{a,j} J_{b,j}^*),$$

the superscript $*$ denotes conjugate transpose, and for any time series of column vectors $\{c_t\}_{t=1}^n$,

$$J_{c,j} = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} c_t \exp(i\omega_j t).$$

Again, the bandwidth m_n satisfies $m_n \to \infty$, $m_n/n \to 0$. They first considered the Type-I stationary case, in which $\widehat{\beta}^{OLS}$ is inconsistent, and showed that

$$\widehat{\beta}_i^{NBLS} - \beta_i = O_p((n/m_n)^{d_U - \widetilde{d}_i})$$

for $i = 1, \ldots, q-1$, where $\widehat{\beta}_i^{NBLS}$ is the i'th entry of $\widehat{\beta}^{NBLS}$, $\widetilde{d}_i$ is the memory parameter of the i'th entry of $\{X_t\}$, and d_U is the memory parameter of $\{U_t\}$. They next considered a Type-II model in which the observable series are nonstationary. Here, the convergence rates (also reported in Marinucci and Robinson 2001) for $\widehat{\beta}_i^{NBLS} - \beta_i$ and $\widehat{\beta}_i^{NBLS} - \beta_i$ are analogous to those obtained for bivariate series in Robinson and Marinucci (2001).

Chen and Hurvich (2006) considered estimation in the Type-I fractional common components model (4), (5), based on the averaged periodogram matrix,

$$I_{m_0} = \sum_{j=1}^{m_0} Re\{J_{y,j}^T J_{y,j}^{T*}\},$$

where m_0 is a fixed positive integer, $J_{y,j}^T$ is the $(q \times 1)$ tapered DFT vector

$$J_{y,j}^T = \sum_{t=1}^{n} h_t^{p-1} y_t \exp(-i\omega_j t),$$

and the taper $\{h_t\}$ is given by (6). Note that I_{m_0}, and statistics based on it, are equivariant to permutation of the entries of $\{y_t\}$. The eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$ of I_{m_0} satisfy $\lambda_j = O_p(n^{2d_k})$, for $j \in N_k$, where $N_0 = \{1, \ldots, a_0\}$ and $N_k = \{(a_0 + \cdots + a_{k-1}) + 1, \ldots, (a_0 + \cdots + a_k)\}$ for $k = 1, \ldots, s$. The eigenvectors corresponding to the j'th largest eigenvalues for $j \in N_k$ converge in distribution to (random) vectors lying in the k'th cointegrating subspace, $k = 0, \ldots, s$. Although these eigenvectors do not consistently estimate fixed population vectors, the estimated space nevertheless converges to the k'th cointegrating subspace in the sense that the norm of the sine of the angle between the true and estimated cointegrating subspaces is $O_p(n^{-\alpha_k})$ where α_k is the shortest gap between the memory parameters corresponding to the given and adjacent subspaces. Gaussian semiparametric estimators constructed from the residuals, *i.e.*, the contemporaneous linear combination of y_t with weights given by an eigenvector in the estimated k'th cointegrating subspace, are asymptotically normal and consistent for d_k, where the bandwidth $m_n \to \infty$, with an upper bound that becomes more restrictive when $\alpha_{min} = \min_k(\alpha_k) = \min(d_0 - d_1, \ldots, d_{s-1} - d_s)$ decreases. If the short-memory component of the spectral density of $\{z_t\}$ is sufficiently smooth, then the upper bound is determined by $(m_n^{2\xi+1}/n^{2\xi}) \log^2 m_n \to 0$, where $\xi = \min(\alpha_{min}, 2)$. If $\alpha_{min} > 1/2$, then the estimator is $m_n^{1/2}$-consistent. If $\alpha_{min} \leq 1/2$ then the estimator is no longer $m_n^{1/2}$-consistent. The restriction on the rate of increase of m_n arises because a linear combination of series with slightly different memory parameters will typically have an irregular short-memory component in its spectral density. Given an *a priori* lower bound on α_{min} it is possible to use the estimates of the d_k to consistently estimate the dimensions $a_0, \ldots, a_s$ of the cointegrating subspaces, as well as s itself and the cointegrating rank $r = a_1 + \cdots + a_s$. This can be accomplished by setting the group boundaries at the points where the sorted estimates of the d_k differ by a sufficient amount. While the need for a lower bound on α_{min} is unfortunate since such a quantity would rarely be known in practice, we note that such lower bounds (assuming $s = 1$) arise implicitly or explicitly in other works on semiparametric fractional cointegration. See Robinson and Yajima (2002), Assumption D, and Velasco (2003), Theorems 2 and 4, as well as Nielsen and Shimotsu (2007).

The estimators for the cointegrating parameter considered above are all direct in the sense that they do not require estimation of memory parameters or other nuisance parameters. An alternative promising approach was proposed by Robinson (2006), in the context of a Type I stationary bivariate system. Under this approach, a bivariate local Whittle estimator is used to jointly estimate the memory parameters $d_{u_1} < d_{u_2} < 1/2$, the parameter β (which is a cointegrating parameter if it is nonzero) and the phase parameter γ in the bivariate system $(y_t, x_t)'$ where $y_t = \beta x_t + u_{1,t}$, $x_t = u_{2,t}$ and the spectral density of $(u_{1,t}, u_{2,t})'$ satisfies (3) with phase parameters $\phi_1 = 0$, $\phi_2 = (\pi/2 - \gamma)(d_{u_2} - d_{u_1})$. If $\beta \neq 0$, then both $\{x_t\}$ and $\{y_t\}$ have memory parameter d_{u_2} but the linear combination $y_t - \beta x_t$ has memory pa-

rameter $d_{u_1} < d_{u_2}$. On the other hand, if $\beta = 0$ there is no cointegration and the memory parameters of $\{x_t\}$ and $\{y_t\}$ are unequal. Assuming sufficient smoothness on the spectral density of $(u_{1,t}, u_{2,t})'$, the resulting local Whittle estimator of β is asymptotically normal, and, to within logarithmic terms, is $n^{2/5+(d_{u_2}-d_{u_1})/5}$-consistent, a faster rate than attained by any of the semiparametric estimators considered above. Furthermore, as $d_{u_2} - d_{u_1}$ approaches 1/2, the local Whittle estimator of β becomes nearly $n^{1/2}$-consistent. The vector of standardized parameter estimates is asymptotically 4-variate normal (the feature of asymptotic normality is not typically shared by the direct estimators of the cointegrating parameter), and the estimate for β converges at a faster rate than the other components. The case $\beta = 0$ is allowed, and tests can be constructed for this hypothesis, which implies no cointegration, but still entails the assumption that $d_{u_1} < d_{u_2}$. Although it is semiparametric, the estimator is nevertheless sensitive to misspecification of the phase parameter. If ϕ_2 is indeed of the form given above, but an incorrect value for γ, say γ^*, is fixed in the local Whittle objective function, which is then minimized with respect to the other parameters, then if $\gamma^* = 0 \bmod (2\pi/(d_{u_2} - d_{u_1}))$ the estimated parameter vector is still asymptotically normal with the same standardization as above, though the limiting covariance matrix becomes more complicated, while if $\gamma^* \neq 0 \bmod (2\pi/(d_{u_2} - d_{u_1}))$ the estimates of d_{u_1} and d_{u_2} are rendered inconsistent and the estimate of β is still asymptotically normal, but with the inferior rate $(n/m_n)^{d_{u_2}-d_{u_1}}$, where $m_n \to \infty$ is the bandwidth used in the estimator. If the form of ϕ_2 were actually other than that assumed, for example $\phi_2 = (\pi/2 - \gamma)(d_{u_2} - d_{u_1})^2$, then all variants considered above of the local Whittle estimator may yield inconsistent estimates of β.

6 Testing for Cointegration; Determination of Cointegrating Rank

Although this chapter has focused mainly on models for fractional cointegration and estimation of the cointegrating parameter, there remains the more basic question of testing for the existence of fractional cointegration. There has been some progress in this direction, which we describe briefly here. A method proposed by Robinson (2006) based on joint local Whittle estimation of all parameters was described in Section 5, although the bivariate model considered there rules out the possibility that the two series have the same memory parameter if there is no cointegration. Marinucci and Robinson (2001) proposed a Hausman-type test in which the memory parameters of the observed series are estimated by two different methods: (1) a multivariate Gaussian semiparametric estimator, imposing the constraint that the memory parameters are the same, and (2) a univariate Gaussian semiparametric estimator of a particular entry. The component of (1) for this entry would

be consistent if and only if there is no cointegration, but the estimator (2) would be consistent in either case. Comparing the standardized differences between these two estimators leads to a test for the null hypothesis H_0 of no cointegration versus the alternative hypothesis of cointegration. Marmol and Velasco (2004), who assumed a nonstationary Type-II model, use regression-based estimators of the cointegrating/projection parameter, and exploit the fact that under H_0 the regressions will be spurious and therefore the true projection parameter is not consistently estimated by OLS. This, together with another estimator of the projection parameter that is consistent under H_0, leads to a Hausman-type test of H_0, under the assumption that if there is cointegration, the cointegrating error is also nonstationary. This assumption would be difficult to maintain if the observed series were volatilities, which are usually considered in the literature to be stationary. Chen and Hurvich (2006) use a statistic based on the difference between the largest and smallest residual-based Gaussian semiparametric estimator of the memory parameter to construct a conservative test of H_0.

If there is indeed cointegration, another question then arises: what is the cointegrating rank? Unlike in the parametric classical cointegration case it is not possible here to handle the problem by multivariate unit root testing as in Johansen (1988, 1991). Robinson and Yajima (2002) proposed a model-selection type method to consistently estimate the cointegrating rank, given an *a priori* lower bound on the degree of memory parameter reduction. A similar method was employed by Chen and Hurvich (2003b), who considered just a single cointegrating subspace, requiring an upper bound on the memory parameter of the cointegrating error and a lower bound on the memory parameter of the observed series. A related method for nonstationary systems based on the exact local Whittle estimator was considered by Nielsen and Shimotsu (2007). A residual-based method for determining not only the cointegrating rank but also the dimensions of the fractional cointegrating subspaces was briefly mentioned in Section 5, and described in more detail in Chen and Hurvich (2006), and in even more detail in a pre-publication version of that paper.

References

Adenstedt, R. K. (1974): On large-sample estimation for the mean of a stationary random sequence. *Annals of Statistics* **2**, 1095–1107.

Akonom, J. and Gourieroux, C. (1987): A functional central limit theorem for fractional processes. Preprint, CEREMAP, Paris.

Breitung, J. and Hassler, U. (2002): Inference on the cointegration rank in fractionally integrated processes. *Journal of Econometrics* **110**, 167–185.

Brockwell, P. J. and R. A. Davis (1991): *Time series: theory and methods: Second Edition.* Springer, New York.

Chan, N. H. and Terrin, N. (1995): Inference for unstable long-memory processes with applications to fractional unit root autoregressions. *Annals of Statistics* **23**, 1662–1683.

Chen, W. (2006): Efficiency in estimation of memory *Preprint, SSRN e-Library*.
Chen, W. and Hurvich, C. (2003a): Estimating fractional cointegration in the presence of polynomial trends. *Journal of Econometrics* **117**, 95–121.
Chen, W. and Hurvich, C. (2003b): Semiparametric estimation of multivariate fractional cointegration. *Journal of the American Statistical Association* **98**, 629–642.
Chen, W. and Hurvich, C. (2006): Semiparametric estimation of fractional cointegrating subspaces. *Annals of Statistics* **34**.
Christensen, B. J. and Nielsen, M. O. (2006): Asymptotic normality of narrow-band least squares in the stationary fractional cointegration model and volatility forecasting. *Journal of Econometrics* **133**, 343–371.
Davidson, J. (2002): A model of fractional cointegration, and tests for cointegration using the bootstrap. *Journal of Econometrics* **110**, 187–212.
Dolado, J.J., Gonzalo, J. and Mayoral, L. (2003): Long range dependence in Spanish political opinion poll data. *Journal of Applied Econometrics* **18**, 137–155.
Dueker, M. and Startz, R. (1998): Maximum-likelihood estimation of fractional cointegration with an application to U.S. and Canadian bond rates. *The Review of Economics and Statistics* **80**, 420–426.
Engle, R. and Granger C. W. J. (1987): Co-integration and error correction: representation, estimation and testing. *Econometrica* **55**, 251–276.
Granger, C. W. J. and Joyeux, R. (1980): An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.
Granger, C. W. J. (1986): Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics* **48**, 213–228.
Hosking, J. R. M. (1981): Fractional differencing. *Biometrika* **68**, 165–176.
Hurvich, C. M. and Chen, W. (2000): An efficient taper for potentially overdifferenced long-memory time series. *Journal of Time Series Analysis* **21**, 155–180.
Hurvich, C. M. and Ray, B. K. (1995): Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis* **16**, 17–41.
Jeganathan, P. (1999): On asymptotic inference in cointegrated time series with fractionally integrated errors. *Econometric Theory* **15**, 583–621.
Johansen, S. (1988): Statistical analysis of cointegration vectors *Journal of Economic Dynamics and Control* **12**, 231–254.
Johansen, S. (1991): Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551–1580.
Lobato, I. N. (1999): A semiparametric two-step estimator in a multivariate long memory model. *Journal of Econometrics* **90**, 129–153.
Mandelbrot, B. and Van Ness, J. (1968): Fractional Brownian motions, fractional noises and applications. *SIAM Review* **10**, 422–437.
Marinucci, D. and Robinson, P.M. (1999): Alternative forms of fractional Brownian motion. *Journal of Statistical Planning and Inference* **80**, 111–122.
Marinucci, D. and Robinson, P.M. (2000): Weak convergence of multivariate fractional processes. *Stochastic Processes and their Applications* **86**, 103–120.
Marinucci, D. and Robinson, P.M. (2001): Journal of Econometrics *Journal of Econometrics* **105**, 225–247.
Marmol, F. and Velasco, C. (2004): Consistent testing of cointegrating relationships. *Econometrica* **72**, 1809–1844.
Nielsen, M. O. and Shimotsu, K. (2007): Determining the cointegrating rank in nonstationary fractional systems by the exact local Whittle approach. Journal of Econometrics to appear.
Pham, T. D., and Tran, L. T. (1985): Some mixing properties of time series models. *Stochastic Processes and their Applications* **19**, 297–303.
Robinson, P. M. (1994): Semiparametric analysis of long-memory time series. *Annals of Statistics* **22**, 515–539.

Robinson, P. M. (1995b): Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* **23**, 1630–1661.

Robinson, P. M. (2006): Multiple local Whittle estimation in nonstationary systems. *Preprint.*

Robinson, P. M. and Hualde, J. (2003): Cointegration in fractional systems with unknown integration orders. *Econometrica* **71**, 1727–1766.

Robinson, P. M. and Marinucci, D. (2001): Narrow-band analysis of nonstationary processes. *Annals of Statistics* **29**, 947–986.

Robinson, P. M. and Marinucci, D. (2003): Semiparametric frequency domain analysis of fractional cointegration. *In: Robinson, P. (Ed.): Time series with long memory, Advanced Texts in Econometrics.* Oxford University Press, Oxford.

Robinson, P. M. and Yajima, Y. (2002): Determination of cointegrating rank in fractional systems. *Econometrics* **106**, 217–241.

Shimotsu, K. (2006): Gaussian semiparametric estimation of multivariate fractionally integrated processes. Journal of Econometrics, forthcoming.

Shimotsu, K. and Phillips, P. C. B. (2005): Exact local Whittle estimation of fractional integration. *Annals of Statistics* **33**, 1890–1933.

Silveira, G. (1991): Contributions to Strong Approximations in time series with applications in nonparametric statistics and functional central limit theorems. *Ph.D. Thesis, University of London.*

Sowell, F. (1990): The fractional unit root distribution. *Econometrica* **58**, 495–505.

Taqqu, M. S. (2003): Fractional Brownian motion and long-range dependence. *In: Doukhan, P., Oppenheim, G. and Taqqu, M. (Eds.): Theory and Applications of Long-Range Dependence.* Birkhäuser, Boston.

Velasco, C. (1999a): Non-stationary log-periodogram regression. *J. Econometrics* **91**, 325–371.

Velasco, C. (1999b): Gaussian semiparametric estimation of non-stationary time series *Journal of Time Series Analysis* **20**, 87–127.

Velasco, C. (2003): Gaussian semi-parametric estimation of fractional cointegration. *Journal of Time Series Analysis* **24**, 345–378.

Velasco, C. and Marmol, F. (2004): Consistent testing of cointegration relationships. *Econometrica* **72**, 1809–1844.

Velasco, C. (2007): The periodogram of fractional processes. Journal of Time Series Analysis, forthcoming.

Different Kinds of Risk

Paul Embrechts, Hansjörg Furrer and Roger Kaufmann

Abstract Over the last twenty years, the financial industry has developed numerous tools for the quantitative measurement of risk. The need for this was mainly due to changing market conditions and regulatory guidelines. In this article we review these processes and summarize the most important risk categories considered.

1 Introduction

Tumbling equity markets, falling real interest rates, an unprecedented increase in longevity, inappropriate reserving, and wrong management decisions were among the driving forces that put the financial stability of so many (insurance) companies at risk over the recent past. Senior management, risk managers, actuaries, accounting conventions, regulatory authorities all played their part. With the solvency of many companies put at stake, political intervention led to the revision of the existing regulatory frameworks. For both the insurance and the banking industry, the aim is to create prudential supervisory frameworks that focus on the true risks being taken by a company.

Paul Embrechts
ETH Zürich, Department of Mathematics, Râmistrasse 101, 8092 Zurich, Switzerland, e-mail: embrechts@math.ethz.ch

Hansjörg Furrer
Swiss Life, General-Guisan-Quai 40, 8022 Zurich, Switzerland, e-mail: Hansjoerg.Furrer@swisslife.ch

Roger Kaufmann
AXA Winterthur, Guisan-Strasse 40, 8401 Winterthur, Switzerland, e-mail: roger.kaufmann@axa.com

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_32,

In the banking regime, these principles were set out by the Basel Committee on Banking Supervision (the "Committee") and culminated in the so-called Basel II Accord, see Basel Committee on Banking Supervision (2005). Initially and under the original 1988 Basel I Accord, the focus has been on techniques to manage and measure market and credit risk. *Market risk* is the risk that the value of the investments will change due to moves in the market risk factors. Typical market risk factors are stock prices or real estate indices, interest rates, foreign exchange rates, commodity prices. *Credit risk*, in essence, is the risk of loss due to counter-party defaulting on a contract. Typically, this applies to bonds where the bond holders are concerned that the counter-party may default on the payments (coupon or principal). The goal of the new Basel II Accord was to overturn the imbalances that prevailed in the original 1988 accord. Concomitant with the arrival of Basel II and its more risk sensitive capital requirements for market and credit risk, the Committee introduced a new risk category aiming at capturing risks "other than market and credit risks". The introduction of the *operational risk* category was motivated, among other considerations, by events such as the Barings Bank failure. The Basel Committee defines operational risk as the risk of loss resulting from inadequate or failed internal processes, people and systems, or from external events. The Basel II definition includes legal risk, but excludes strategic risk, i.e. the risk of a loss arising from a poor strategic business decision. Furthermore, this definition excludes reputational risk. Examples of operational risk include, among others, technology failure, business premises becoming unavailable, errors in data processing, fraud, etc. The capital requirement of Basel II is that banks must hold capital of at least 8% of total risk-weighted assets. This definition was retained from the original accord.

Insurance regulation too is rapidly moving towards risk-based foundations. Based on the findings of the Müller Report Müller (1997), it was recognized that a fundamental review of the assessment of the overall financial position of an insurance company should be done, including for example the interactions between assets and liabilities, accounting systems and the methods to calculate the solvency margins. In 2001, the European Commission launched the so-called Solvency II project. The key objective of Solvency II is to secure the benefits of the policyholders thereby assessing the company's overall risk profile. A prudential supervisory scheme does not strive for a "zero-failure" target; in a free market, failures will occur. Rather, prudential supervisory frameworks should be designed in such a way that a smooth run-off of the portfolios is ensured in case of financial distress. Phase 1 of the Solvency II project began in 2001 with the constitution of the so-called London Working Group chaired by Paul Sharma from the FSA (Financial Services Authority). The resulting Sharma Report Sharma (2002) was published in 2002, and contains a survey of actual failures and near misses from 1996 to 2001. The second phase lasts from 2003 to 2007 and is designated to the development of more detailed rules. Finally, the third phase should be terminated

by 2010, and is devoted to the implementation of the new standards, also in the national laws.

At the heart of both Basel II and Solvency II lies a *three pillar structure*. Pillar one defines the minimum financial requirements. The second pillar earmarks the supervisory review process, whereas pillar three sets out the disclosure requirements. The minimum financial requirements relate a company's available capital to its economic capital. Economic capital is the amount of capital that is needed to support for retained risks in a loss situation. Associating available capital with economic capital is only meaningful if consistency prevails between valuation and risk measurement. The arrival of a robust marked-to-market culture in the 1990s helps to achieve greater harmonization in this context.

The three-pillar structure of both risk based insurance and banking supervisory frameworks indicates that the overall assessment of a financial institution's financial stability goes beyond the determination of capital adequacy ratios. Nevertheless, the focus in this note will be on the capital requirements, that is on pillar one. More specifically, we address the issue of how to quantify market, credit and insurance risk. We also touch upon the measurement of operational risk. But rather than promoting seemingly sophisticated (actuarial) measurement techniques for quantifying operational risk, we focus on the very special nature of this risk category, implying that standard analytical concepts prove insufficient and also yield counter-intuitive results in terms of diversification.

In hindsight, the inexperienced reader could be tempted to believe that only regulators demand for distinctive risk management cultures and cutting-edge economic capital models. Alas, there are many more institutions that keep a beady eye on the companies' risk management departments: analysts, investors, and rating agencies, to name a few, have a growing interest in what is going on on the risk management side. Standard and Poor's for instance, a rating agency, recently added an "Enterprise Risk Management" (ERM) criterion when rating insurance companies. The ERM rating is based on five key metrics, among which are the risk and economic capital models of insurance undertakings.

The remainder of this note is organized as follows. In Section 2 we provide the basic prerequisites for quantitative risk management by introducing the notion of risk measures and the concept of risk factor mapping. Special emphasis will be given to two widely used risk measures, namely Value at Risk (VaR) and expected shortfall. Section 3 is devoted to the measurement of credit risk, whereas Section 4 deals with market risk. The problem of how to scale a short term VaR to a longer term VaR will be addressed in Section 4.3. The particularities of operational risk loss data and their implications on the economic capital modeling in connection with VaR will be discussed in Section 5. Section 6 is devoted to the measurement of insurance risk. Both the life and non-life measurement approach that will be presented originate from the Swiss Solvency Test. In Section 7 we make some general comments on

the aggregation of risks in the realm of economic capital modeling. Attention will be drawn to risks that exhibit special properties such as extreme heavy-tailedness, extreme skewness, or a particular dependence structure.

2 Preliminaries

Risk models typically aim at quantifying likely losses of a portfolio over a given time horizon that could incur for a variety of risks. Formal risk modeling for instance is required under the new (risk-sensitive) supervisory frameworks in the banking and insurance world (Basel II, Solvency II). In this section, we provide the prerequisites for the modeling of risk by introducing risk measures and the notion of risk factor mapping.

2.1 Risk measures

The central notion in actuarial and financial mathematics is the notion of uncertainty or risk. In this article, uncertainty or risk will always be represented by a random variable, say X or $X(t)$, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T^*]}, \mathbb{P})$. The filtration $(\mathcal{F}_t)_t$ is assumed to satisfy the "usual conditions", that is a) $(\mathcal{F}_t)_t$ is right-continuous and b) $\mathcal{F}_0$ contains all null sets, i.e. if $B \subset A \in \mathcal{F}_0$ with $\mathbb{P}[A] = 0$, then $B \in \mathcal{F}_0$.

Since risks are modeled as (non-negative) random variables, measuring risk is equivalent to establishing a relation ϱ between the set of random variables and $\mathbb{R}$, the real numbers. Put another way, a risk measure is a function mapping a risk X to a real number $\varrho(X)$. If for example X defines the loss in a financial portfolio over some time horizon, then $\varrho(X)$ can be interpreted as the additional amount of capital that should be set aside so that the portfolio becomes acceptable for a regulator, say. The definition of a risk measure is very general, and yet risk measures should fulfill certain properties to make them "good" risk measures. For instance, it should always hold that $\varrho(X)$ is bounded by the largest possible loss, as modeled by F_X. Within finance, Artzner et al. (1999) pioneered the systematic study of risk measure properties, and defined the class of so-called "coherent risk measures" to be the ones satisfying the following properties:

(a) Translation invariance: $\varrho(X+c) = c+\varrho(X)$, for each risk X and constant $c > 0$.
(b) Positive homogeneity: $\varrho(cX) = c\varrho(X)$, for all risks X and constants $c > 0$.
(c) Monotonicity: if $X \leq Y$ a.s., then $\varrho(X) \leq \varrho(Y)$.
(d) Subadditivity: $\varrho(X + Y) \leq \varrho(X) + \varrho(Y)$.

Subadditivity can be interpreted in the way that "a merger should not create extra risk"; it reflects the idea that risk in general can be reduced via diversification.

Note that there exist several variations of these axioms depending on whether or not losses correspond to positive or negative values, or whether a discount rate over the holding period is taken into account. In our case, we consider losses as positive and neglect interest payments. The random variables X, Y correspond to values of risky positions at the end of the holding period, hence the randomness.

2.1.1 Value at Risk

The most prominent risk measure undoubtedly is *Value at Risk* (VaR). It refers to the question of how much a portfolio position can fall in value over a certain time period with a given probability. The concept of Value at Risk originates from J.P. Morgan's RiskMetrics published in 1993. Today, VaR is the key concept in the banking industry for determining market risk capital charges. A textbook treatment of VaR and its properties is Jorion (2000). Formally, VaR is defined as follows:

Definition 1 Given a risk X with cumulative distribution function F_X and a probability level $\alpha \in (0, 1)$, then

$$\mathrm{VaR}_\alpha(X) = F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}.$$

Typical values for α are 0.95, 0.99 or 0.999. The Basel II approach for a market risk charge for example requires a holding period of ten days and a confidence level of $\alpha = 0.99$. At the trading floor level, individual trading limits are typically set for one day, $\alpha = 0.95$.

Even though VaR has become the benchmark risk measure in the financial world, it has some deficiencies which we shall address here. First, note that VaR only considers the result at the end of the holding period, hence it neglects what happens with the portfolio value along the way. Moreover, VaR assumes the current positions being fixed over the holding period. In practice, however, positions are changed almost continuously. It is fair to say, however, that these weaknesses are not peculiar to VaR; other one-period risk measures have the same shortcomings. More serious though is the fact that VaR does *not* measure the potential size of a loss given that the loss exceeds VaR. It is mainly for this reason why VaR is not being used in the Swiss Solvency Test for the determination of the so-called target capital. There, the regulatory capital requirement asks for sufficient capital to be left (on average) in a situation of financial distress in order to ensure a smooth run-off of the portfolio.

The main criticism of VaR, however, is that in general it lacks the property of subadditivity. Care has to be taken when risks are extremely skewed or

heavy-tailed, or in case they encounter a special dependency structure. In such circumstances, VaR may not be sub-additive, as the following example with two very heavy-tailed risks shows. The implications for the modeling of economic capital are severe as the concept of diversification breaks down. We come back to this issue later in Sections 5 and 7 when we talk about operational risk losses and their aggregation.

Example 1 Let X_1 and X_2 be two independent random variables with common distribution function $F_X(x) = 1-1/\sqrt{x}$ for $x \geq 1$. Observe that the risks X_1, X_2 have infinite mean, and thus are very heavy-tailed. Furthermore, one easily shows that $\mathrm{VaR}_\alpha(X) = (1-\alpha)^{-2}$. Straightforward calculation then yields

$$\begin{aligned} F_{X_1+X_2}(x) &= \mathbb{P}[X_1 + X_2 \leq x] \\ &= \int_1^{x-1} F_X(x-y)\, dF_X(y) \\ &= 1 - 2\sqrt{x-1}/x \\ &< 1 - \sqrt{2/x} \\ &= F_{2X}(x), \end{aligned}$$

where $F_{2X}(u) = \mathbb{P}[2X_1 \leq u]$ for $u \geq 2$. From this, we then conclude that $\mathrm{VaR}_\alpha(X_1 + X_2) > \mathrm{VaR}_\alpha(2X_1)$. Since $\mathrm{VaR}_\alpha(2X_1) = \mathrm{VaR}_\alpha(X_1) + \mathrm{VaR}_\alpha(X_1)$, it follows that

$$\mathrm{VaR}_\alpha(X_1 + X_2) > \mathrm{VaR}_\alpha(X_1) + \mathrm{VaR}_\alpha(X_2),$$

hence demonstrating that VaR is not sub-additive in this case. Note that a change in the risk measure from VaR to expected shortfall (see Definition 2 below), say, is no reasonable way out in this case. The problem being that expected shortfall is infinite in an infinite mean model.

We conclude this section by showing that VaR is sub-additive for normally distributed risks. In fact, one can show that VaR is sub-additive for the wider class of linear combinations of the components of a multivariate elliptical distribution, see for instance McNeil et al. (2005), Theorem 6.8.

Example 2 Let X_1, X_2 be jointly normally distributed with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where $-1 \leq \rho \leq 1$ and $\sigma_i > 0$, $i = 1, 2$. Let $0.5 \leq \alpha < 1$, then

$$\mathrm{VaR}_\alpha(X_1 + X_2) \leq \mathrm{VaR}_\alpha(X_1) + \mathrm{VaR}_\alpha(X_2). \tag{1}$$

The main observation here is that, since (X_1, X_2) is bivariate normally distributed, X_1, X_2 and X_1+X_2 all have univariate normal distributions. Hence it follows that

$$\mathrm{VaR}_\alpha(X_i) = \mu_i + \sigma_i\, q_\alpha(N), \qquad i = 1, 2,$$
$$\mathrm{VaR}_\alpha(X_1 + X_2) = \mu_1 + \mu_2 + \sqrt{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}\; q_\alpha(N).$$

Here $q_\alpha(N)$ denotes the α-quantile of a standard normally distributed random variable. The assertion in (1) now follows because of $q_\alpha(N) \geq 0$ (since $0.5 \leq \alpha < 1$) and $(\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2)^{1/2} \leq \sigma_1 + \sigma_2$ (since $\rho \leq 1$).

2.1.2 Expected shortfall

As mentioned earlier, for a given level α, VaR_α does not give information on the loss sizes beyond this quantile. To circumvent this problem, Artzner et al. (1999) considered the notion of expected shortfall or conditional tail expectation instead.

Definition 2 Let X be a risk and $\alpha \in (0, 1)$. The expected shortfall or conditional tail expectation is defined as the conditional expected loss given that the loss exceeds $\mathrm{VaR}_\alpha(X)$:

$$\mathrm{ES}_\alpha(X) = \mathbb{E}[X | X > \mathrm{VaR}_\alpha(X)]\,.$$

Intuitively, $\mathrm{ES}_\alpha(X)$ represents the average loss in the worst $100(1-\alpha)\%$ cases. This representation is made more precise by observing that for a continuous random variable X one has

$$\mathrm{ES}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\xi(X)\, d\xi, \qquad 0 < \alpha < 1.$$

For continuous risks X expected shortfall, as defined in Definition 2, is a coherent risk measure, see Artzner et al. (1999). For risks which are not continuous, a slight modification of Definition 2 leads to a coherent, i.e. subadditive risk measure; see McNeil et al. (2005), Section 2.2.4.

2.2 Risk factor mapping and loss portfolios

Denote the value of a portfolio at time t by $V(t)$. The loss of the portfolio over the period $[t, t+h]$ is given by

$$L_{[t,t+h]} = -\big(V(t+h) - V(t)\big).$$

Note our convention to quote losses as positive values. Following standard risk management practice, the portfolio value is modeled as a function of a d-dimensional random vector $\mathbf{Z}(t) = (Z_1(t), Z_2(t), \ldots, Z_d(t))'$ of risk factors Z_i. Hence,

$$V(t) = V(t; \mathbf{Z}(t)). \tag{2}$$

The representation (2) is known as mapping of risks. Representing a financial institution's portfolio as a function of underlying market-risky instruments constitutes a crucial step in any reasonable risk management system. Indeed, any potential risk factor *not* included in this mapping will leave a blind spot on the resulting risk map.

It is convenient to introduce the vector $\mathbf{X}_{[t,t+h]} = \mathbf{Z}(t+h) - \mathbf{Z}(t)$ of risk factor changes for the portfolio loss $L_{[t,t+h]}$. It can be approximated by $L^{\Delta}_{[t,t+h]}$, where

$$L^{\Delta}_{[t,t+h]} = (\nabla V)' \mathbf{X}_{[t,t+h]} \tag{3}$$

provided the function $V : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Here, ∇f denotes the vector of partial derivatives $\nabla f = (\partial f/\partial z_1, \ldots, \partial f/\partial z_d)'$. Observe that in (3) we suppressed the explicit time dependency of V. The approximation (3) is convenient as it allows one to represent the portfolio loss as a linear function of the risk factor changes, see also Section 4.1.1. The linearity assumption can be viewed as a first-order approximation (Taylor series expansion of order one) of the risk factor mapping. Obviously, the smaller the risk factor changes, the better the quality of the approximation.

3 Credit Risk

Credit risk is the risk of default or change in the credit quality of issuers of securities to whom a company has an exposure. More precisely, default risk is the risk of loss due to a counter-party defaulting on a contract. Traditionally, this applies to bonds where debt holders are concerned that the counter-party might default. Rating migration risk is the risk resulting from changes in future default probabilities. For the modeling of credit risk, the following elements are therefore crucial:

- *default probabilities*: probability that the debtor will default on its obligations to repay its debt;
- *recovery rate*: proportion of the debt's par value that the creditor would receive on a defaulted credit, and
- *transition probabilities*: probability of moving from one credit quality to another within a given time horizon.

In essence, there are two main approaches for the modeling of credit risk, so-called *structural models* and *reduced form* or *intensity based* methods.

3.1 Structural models

Merton (1974) proposed a simple capital structure of a firm where the dynamics of the assets are governed by a geometric Brownian motion:

$$dA(t) = A(t)(\mu\, dt + \sigma\, dW(t)), \qquad t \in [0, T].$$

In its simplest form, an obligor's default in a structural model is said to occur if the obligor's asset value $A(T)$ at time T is below a pre-specified deterministic barrier x, say. The default probability can then be calculated explicitly:

$$\mathbb{P}[A(T) \leq x] = \Phi\left(\frac{\log(x/A(0)) - (\mu - \sigma^2/2)T}{\sigma\sqrt{T}}\right).$$

Here Φ denotes the cumulative distribution function of a standard normal random variable, i.e. $\Phi(x) = 1/\sqrt{2\pi}\int_{-\infty}^{x} \exp\{-y^2/2\}\, dy$. Various extensions of Merton's original firm value model exist. For instance, one can let the barrier x be a (random) function of time.

3.2 Reduced form models

In a reduced form pricing framework, it is assumed that the default time τ is governed by a risk neutral default intensity process $\lambda = \{\lambda(t) : t \geq 0\}$. That is, default is defined as the first arrival time (jump time) τ of a counting process with intensity λ. It can be shown that the conditional probability at time t, given all information at that time, of survival to a future time T, is given by

$$p(t, T) = \mathbb{E}_{\mathbb{Q}}\left[e^{-\int_t^T \lambda(u)\, du} | \mathcal{F}_t\right] \tag{4}$$

From (4) one immediately recognizes the analogy between an intensity process λ and a short interest rate process r for the time-t price of a (default free) zero-coupon bond maturing at time T. The latter is given by

$$P(t, T) = \mathbb{E}_{\mathbb{Q}}\left[B(t)/B(T) | \mathcal{F}_t\right] = \mathbb{E}_{\mathbb{Q}}\left[e^{-\int_t^T r(u)\, du} | \mathcal{F}_t\right],$$

where $B(t) = \exp\{\int_0^t r(s)\, ds\}$ denotes the risk free bank account numéraire. As shown by Lando (1998), the *defaultable* bond price at time t (assuming zero recovery) is then given by

$$\bar{P}(t, T) = \mathbb{E}_{\mathbb{Q}}\left[e^{-\int_t^T (r(u) + \lambda(u))\, du} | \mathcal{F}_t\right],$$

provided default has not already occurred by time t.

Reduced form models can be extended to allow for non-zero recovery. Duffie and Singleton (1999) for instance introduce the concept of recovery of market value (RMV), where recovery is expressed as a fraction of the market value of the security just prior to default. Formally, it is assumed that the claim pays $(1 - L(t))V(t-)$, where $V(t-) = \lim_{s \uparrow t} V(s)$ is the price of the claim just before default, and $L(t)$ is the random variable describing the fractional loss of market value of the claim at default. Under technical conditions, Duffie and Singleton (1999) show that

$$\bar{P}(t,T) = \mathbb{E}_{\mathbb{Q}}\left[e^{-\int_t^T (r(u)+\lambda(u)L(u))\,du} | \mathcal{F}_t\right].$$

Here, $r(u) + \lambda(u)L(u)$ is the default-adjusted short rate.

3.3 Credit risk for regulatory reporting

Compared to the original 1988 Basel accord and its amendments, Basel II better reflects the relative credit qualities of obligors based on their credit ratings. Two approaches are being proposed under Basel II, namely

(A) Standardized approach
(B) Internal-ratings-based approach

The *standardized approach* better recognizes the benefits of credit risk migration and also allows for a wider range of acceptable collateral. Under the *internal-ratings-based approach*, a bank can — subject to the bank's regulator approval — use its own internal credit ratings. The ratings must correspond to the one-year default probabilities and have to be in place for a minimum of three years.

The assessment of credit risk under the Solvency II regime essentially follows the Basel II principles. Within the Swiss Solvency Test for instance, the Basel II standardized approach is being advocated. Portfolio models are acceptable too, provided they capture the credit migration risk. It is for this reason why the CreditRisk+ model for instance would not be permissible within the Swiss Solvency Test as this model only covers the default risk, but not the credit migration risk.

4 Market Risk

Market risk is the risk that the value of an investment will decrease due to moves in market risk factors. Standard market risk factors are interest rates, stock indices, commodity prices, foreign exchange rats, real estate indices, etc.

4.1 Market risk models

4.1.1 Variance-covariance

The standard analytical approach to estimate VaR or expected shortfall is known as *variance-covariance method*. This means that the risk factor changes are assumed to be samples from a multivariate normal distribution, and that the loss is represented as a linear function of the risk factor changes, see Section 2.2 for more details. This approach offers an analytical solution, and it is much faster to calculate the risk measures in a parametric regime than performing a simulation. However, a parametric approach has significant limitations. The assumption of normally distributed risk factor changes may heavily underestimate the severeness of the loss distribution. Moreover, linearization may be a poor approximation of the risk factor mapping.

4.1.2 Historical simulation

The second well-established approach for measuring market risk exposure is the *historical simulation method*. Instead of estimating the loss distribution under some explicit parametric model for the risk factor changes, one uses the empirical distribution of the historical loss data. VaR and expected shortfall can then either be estimated directly from the simulated data or by first fitting a univariate distribution to the loss data. The main advantage of this approach is its simplicity of implementation. No statistical estimation of the distribution of the risk factor changes is required. In particular, no assumption on the interdependence of the underlying risk factors is made. On the downside, it may be difficult to collect enough historical data of good quality. Also, the observation period is typically not long enough such that samples of extreme changes in the portfolio value cannot be found. Therefore, adding suitable stress scenarios is very important.

4.1.3 Monte Carlo simulation

The idea behind the *Monte Carlo method* is to estimate the loss distribution under some explicit parametric model for the risk factor changes. To be more precise, one first fits a statistical model to the risk factor changes. Typically, this model is inferred from the observed historical data. Monte Carlo-simulated risk factor changes then allow one to make inferences about the loss distribution and the associated risk measure. This approach is very general, albeit it may require extensive simulation.

4.2 Conditional versus unconditional modeling

In an *unconditional approach*, one neglects the evolution of risk factor changes up to the present time. Consequently, tomorrow's risk factor changes are assumed to have the same distribution as yesterday's, and thus the same variance as experienced historically. Such a stationary model corresponds to a long-term view and is often appropriate for insurance risk management purposes. However, empirical analysis often reveals that the volatility σ_t of market risk factor changes $X(t)$, *conditionally* on their past, fluctuates over time. Sometimes, the market is relatively calm, then a crisis happens, and the volatility will suddenly increase. Time series models such as GARCH type models allow the variance σ_{t+1}^2 to vary through time. They are suited for a short-term perspective. GARCH stands for generalized autoregressive conditional heteroskedasticity, which in essence means that the conditional variance on one day is a function of the conditional variances on the previous days.

4.3 Scaling of market risks

A risk measure's holding period should be related to the liquidity of the assets. If a financial institution runs into difficulties, the holding period should cover the time necessary to raise additional funds for corrective actions. The Basel II VaR approach for market risk for instance requires a holding period of ten days (and a confidence level $\alpha = 0.99$). The time horizon thus often spans several days, and sometimes even extends to a whole year. While the measurement of short-term financial risks is well established and documented in the financial literature, much less has been done in the realm of long-term risk measurement. The main problem is that long-dated historical data in general is not representative for today's situation and therefore should not be used to make forecasts about the future changes in market risk factors. So the risk manager is typically left with little reliable data to make inference of long term risk measures.

One possibility to close this gap is to *scale* a short-term risk estimate to a longer one. The simplest way to do this is to use the square-root-of-time scaling rule, where a k-day Value at Risk $\mathrm{VaR}^{(k)}$ is scaled with $\sqrt{n}$ in order to get an estimate for the nk-day Value at Risk $\mathrm{VaR}^{(nk)} \approx \sqrt{n}\mathrm{VaR}^{(k)}$. This rule is motivated by the fact that one often considers the *logarithm* of tradable securities, say S, as risk factors. The return over a 10-day period for example is then expressible as $R_{[0,10]} := \log(S(10))/S(0)) = X(1) + X(2) + \cdots + X(10)$, where $X(k) = \log(S(k)) - \log(S(k-1)) = \log(S(k)/S(k-1))$, $k = 1, 2, \ldots, 10$. Observe that for independent random variables $X(i)$ the standard deviation of $R_{[0,10]}$ equals $\sqrt{10}$ times the standard deviation of $X(1)$. In this section we analyze under which conditions such scaling is appropriate.

We concentrate on unconditional risk estimates and on VaR as risk measure. Recall our convention to quote losses as positive values. Thus the random variable $L(t)$ will subsequently denote the negative value of one-day log-returns, i.e. $L(t) = -\log(S(t)/S(t-1))$ for some security S.

4.3.1 Scaling under normality

Under the assumption of independent and identically zero-mean normally distributed losses $L(t) \sim \mathcal{N}(0, \sigma^2)$ it follows that the n-day losses are also normally distributed, that is $\sum_{t=1}^{n} L(t) \sim \mathcal{N}(0, n\sigma^2)$. Recall that for a $\mathcal{N}(0, \widetilde{\sigma}^2)$-distributed loss L, VaR is given by $\mathrm{VaR}_\alpha(L) = \widetilde{\sigma}\, q_\alpha(N)$, where $q_\alpha(N)$ denotes the α-quantile of a standard normally distributed variate. Hence the square-root-of-time scaling rule

$$\mathrm{VaR}^{(n)} = \sqrt{n}\,\mathrm{VaR}^{(1)}$$

works perfectly in this case.

Now let a constant value μ be added to the one-day returns, i.e. μ is subtracted from the one-day loss: $L(t) \sim \mathcal{N}(-\mu, \sigma^2)$. Assuming independence among the one-day losses, the n-day losses are again normally distributed with mean value $-n\mu$ and variance $n\sigma^2$, hence $\sum_{t=1}^{n} L(t) \sim \mathcal{N}(-n\mu, n\sigma^2)$. The VaR in this case will be increased by the trend of L. This follows from $\mathrm{VaR}^{(n)} + n\mu = \sqrt{n}\,(\mathrm{VaR}^{(1)} + \mu)$, or equivalently

$$\mathrm{VaR}^{(n)} = \sqrt{n}\,\mathrm{VaR}^{(1)} - (n - \sqrt{n})\mu.$$

Accounting for trends is important and therefore trends should never be neglected in a financial model. Note that the effect increases linearly with the length n of the time period.

To simplify matters, all the models presented below are restricted to the zero-mean case. They can easily be generalized to non-zero mean models, implying that the term $(n-\sqrt{n})\mu$ must be taken into account when estimating and scaling VaR.

4.3.2 Autoregressive models

Next, we consider a stationary autoregressive model of the form

$$L(t) = \lambda L(t-1) + \varepsilon_t,$$

where $(\varepsilon_t)_{t\in\mathbb{N}}$ is a sequence of iid zero-mean normal random variables with variance σ^2 and $\lambda \in (-1, 1)$. Not only are the one-day losses normally distributed, but also the n-day losses:

$$L(t) \sim \mathcal{N}\big(0, \frac{\sigma^2}{1-\lambda^2}\big), \qquad \sum_{t=1}^{n} L(t) \sim \mathcal{N}\big(0, \frac{\sigma^2}{(1-\lambda)^2}(n - 2\lambda\frac{1-\lambda^n}{1-\lambda^2})\big)\,.$$

Hence, making use of $\mathrm{VaR}_\alpha(L) = \widetilde{\sigma}\, q_\alpha(N)$, one obtains

$$\mathrm{VaR}^{(n)} = \sqrt{\frac{1+\lambda}{1-\lambda}\left(n - 2\lambda\frac{1-\lambda^n}{1-\lambda^2}\right)}\ \mathrm{VaR}^{(1)}. \tag{5}$$

Since the square-root expression in (5) tends to $\sqrt{n}$ as $\lambda \to 0$, one concludes that the scaled one-day value $\sqrt{n}\,\mathrm{VaR}^{(1)}$ is a good approximation of $\mathrm{VaR}^{(n)}$ for small values of λ.

For more general models, such as stochastic volatility models with jumps or AR(1)-GARCH(1,1) models, the correct scaling from a short to a longer time horizon depends on the confidence level α and cannot be calculated analytically. In many practical applications, the confidence level varies from 0.95 to 0.99, say. Empirical studies show that for such values of α, scaling a short-term VaR with the square-root-of-time yields a good approximation of a longer-term VaR, see Kaufmann (2005). For smaller values of α, however, the scaled risks tend to *overestimate* the true risks, whereas larger values of α tend to *underestimate* the risks. In the limit $\alpha \to 1$, one should abstain from scaling risks, see Brummelhuis and Guégan (2000, 2005).

Sometimes risk managers are also confronted with the problem of transforming a 1-day VaR at the confidence level $\alpha = 0.95$ to a 10-day VaR at the 0.99 level. From a statistical viewpoint, such scaling should be avoided. Our recommendation is to first try to arrive at an estimate of the 1-day VaR at the 0.99 level and then to make inference of the 10-day VaR by means of scaling.

In this section we analyzed the scaling properties in a VaR context. As a matter of fact, these properties in general do not carry over when replacing VaR through other risk measures such as expected shortfall. In an expected shortfall regime coupled with heavy-tailed risks, scaling turns out to be delicate. For light-tailed risk though the square-root-of-time rule still provides good results when expected shortfall is being used.

5 Operational Risk

According to the capital adequacy frameworks as set out by the Basel Committee, the general requirement for banks is to hold total capital equivalent to at least 8% of their risk-weighted assets. This definition was retained of the old capital adequacy framework (Basel I). In developing the revised framework now known as Basel II the idea was to arrive at significantly more risk-sensitive capital requirements. A key innovation in this regard was that

operational risk –besides market and credit risk– must be included in the calculation of the total minimum capital requirements. Following the Committee's wording, we understand by operational risk "the risk of losses resulting from inadequate or failed internal processes, people and systems, or external events".

The Basel II framework provides a range of options for the determination of an operational risk capital charge. The proposed methods allow banks and supervisors to select approaches that are most appropriate for a bank's operations. These methods are:

(1) basic indicator approach,
(2) standardized approach,
(3) advanced measurement approach.

Both the *basic indicator approach* as well as the *standardized approach* are essentially volume-based measurement methods. The proxy in both cases is the average gross income over the past three years. These measurement methods are primarily destined for small and medium-sized banks whose exposure to operational risk losses is deemed to be moderate. Large internationally active banks, on the other hand, are expected to implement over time a more sophisticated measurement approach. Those banks must demonstrate that their approach is able to capture "severe tail loss events". More formally, banks should set aside a capital charge C_{Op} for operational risk in line with the 99.9% confidence level on a one-year holding period. Using VaR as risk measure, this approach is known as *loss distribution approach* (LDA). It is suggested to use $\sum_{k=1}^{8} \mathrm{VaR}_\alpha(L_k)$ for a capital charge and to allow for a capital reduction through diversification under appropriate dependency assumptions. Here, L_k denotes the one-year operational risk loss of business line k. The choice of 8 business lines and their precise definition is to be found in the Basel II Accord, banks are allowed to use fewer or more.

It is at this point where one has to sway a warning flag. A recent study conducted by Moscadelli (2004) reveals that operational loss amounts are *very* heavy-tailed. This stylized fact has been known before, albeit not in such an unprecedented way. Moscadelli's analysis suggests that the loss data from six out of eight business lines come from an infinite mean model! An immediate consequence is that standard correlation coefficients between two such one-year losses do not exist. Nešlehová et al. (2006) in their essay carry on with the study of Moscadelli's data and show the serious implications extreme heavy-tailedness can have on the economic capital modeling, in particular when using VaR as a risk measure. Note that it is not the determination of a VaR per se that causes problems in an infinite mean model. Rather, it is the idea of capital reduction due to aggregation or pooling of risks that breaks down in this case, see Example 1 on page 734. We will come back to this issue later in Section 7.

Operational risk is also part of Solvency II and most of the insurance industry's national risk-based standard models. In the realm of the Swiss

Solvency Test for instance it suffices to assess operational risk on a pure qualitative basis. Other models, such as the German GDV model for instance, require a capital charge for operational risk. This charge is mainly volume-based, similar to the Basel II basic indicator or standardized approach.

6 Insurance Risk

6.1 Life insurance risk

Life insurance contracts are typically characterized by long-term financial promises and guarantees towards the policyholders. The actuary's main task has therefore been to forecast the future liabilities, that is to set up sufficient reserves in order that the company can meet its future obligations. Ideally, the modern actuary should also be able to form an opinion on how many assets will be required to meet the obligations and on how the asset allocation should look like from a so-called *asset and liability management* (ALM) perspective. So life insurance companies are heavily exposed to *reserve* risk. Under reserve risk, we understand the risk that the actual claims experience deviates from the booked reserves. Booked reserves are always based on some accounting conventions and are determined in such a way that sufficient provisions are held to cover the *expected* actuarial liabilities based on the tariffs. Typically, these reserves are formula-based, that is, a specific calculation applied individually to each contract in force, then summed up, yields the reserves. Even though they include a margin for prudence, the reserves may prove insufficient in the course of time because of e.g. demographic changes.

Reserve risk can further be decomposed into the following sub-categories:

(A) stochastic risk,
(B) parametric risk,
(C) model risk.

The *stochastic risk* is due to the variation and severity of the claims. In principle, the stochastic risk can be diversified through a greater portfolio and an appropriate reinsurance program. By ceding large individual risks to a reinsurer via a surplus share for instance, the portfolio becomes aptly homogeneous.

Parametric risk arises from the fact that tariffs can be subject to material changes over time. For example, an unprecedented increase in longevity implies that people will draw annuities over a longer period. It is the responsibility of the (chief) actuary to continually assess and monitor the adequacy of the reserves. Periodic updates of experience data give insight into the adequacy of the reserves.

By *model risk* finally we understand the risk that a life insurance company has unsuitable reserve models in place. This can easily be the case

when life insurance products encompass a variety of policyholder options such as e.g. early surrender or annuity take-up options. Changing economic variables and/or an increase in the longevity can result in significant future liabilities, even when the options were far out of the money at policy inception. It was a combination of falling long-term interest rates and booming stock markets coupled with an increase in longevity that put the solvency of Equitable Life, a UK insurer, at stake and led to the closure of new business. The reason for this was that so-called guaranteed annuity options dramatically increased in value and subsequently constituted a significant liability which was neither priced nor reserved for. Traditional actuarial pricing and reserving methods based on the expectation pricing principle prove useless in this context were it not for those policyholders who behave in a financially irrational way. Indeed, empirical studies may reveal that there is *no* statistical evidence supporting a link between the surrender behavior and the level of market interest rates. Arbitrage pricing techniques are always based on the assumption of financially rational policyholder behavior though. This means that a person would surrender its endowment policy at the first instant when the actual payoff exceeded the value of continuation.

The merits of arbitrage pricing techniques are that they provide insight into the mechanism of embedded options, and consequently these findings should be used when designing new products. This will leave an insurance company immune against potential future changes in the policyholders' behavior towards a more rational one from a mathematical economics point of view.

6.2 Modeling parametric life insurance risk

In the following, we will present a model that allows for the quantification of parametric life insurance risk. This model is being used within the Swiss Solvency Test. In essence, it is a variance-covariance type model, that is

- risk factor changes have a multivariate normal distribution, and
- changes in the best estimate value of liabilities linearly depend on the risk factor changes.

More formally, it is assumed that for risk factor changes $\mathbf{X}$ and weights b

$$\Delta L = \mathbf{b}'\mathbf{X}$$

where $L = L(\mathbf{Z}(t))$ denotes the best estimate value of liabilities at the valuation date t, and $\mathbf{Z}(t) = \big(Z_1(t), Z_2(t), \ldots, Z_d(t)\big)'$ is the vector of (underwriting) risk factors. Best estimate values are unbiased (neither optimistic, nor pessimistic, nor conservative) estimates which employ the most recent and accurate actuarial and financial market information. Best estimate values are

without any (safety) margins whatsoever. Typically, the value L is obtained by means of projecting the future cash flows and subsequent discounting with the current risk-free yield curve.

Again, we denote the risk factor changes by $\mathbf{X}(t) = \mathbf{Z}(t) - \mathbf{Z}(t-1)$. Within the Swiss Solvency Test, the following set of risk factors is considered:

Table 1 Life insurance risk factors in the Swiss Solvency Test.

(R1)	mortality	(R4)	recovery
(R2)	longevity	(R5)	surrender/lapse
(R3)	disability	(R6)	annuity take-up

The risk factor "mortality" for example refers to the best estimate one-year mortality rates q_x, q_y respectively (second-order mortality rates). The risk factor "longevity" refers to the improvement of mortality which is commonly expressed in exponential form

$$q(x,t) = q(x,t_0)e^{-\lambda_x(t-t_0)}, \qquad t \geq t_0,$$

where $q(x,t_0)$ stands for the best estimate mortality rate of an x year old male at time t_0.

Typically, no analytical solutions exist for the partial derivatives $b_k = \partial L/\partial z_k$, and hence they have to be approximated numerically by means of sensitivity calculations:

$$b_k \approx \frac{L(\mathbf{Z} + \varepsilon \mathbf{e}_k) - L(\mathbf{Z})}{\varepsilon}$$

for ε small, e.g. $\varepsilon = 0.1$. Here, $\mathbf{e}_k = (0, \ldots, 0, 1, 0, \ldots, 0)'$ denotes the kth basis vector in $\mathbb{R}^d$. Combining everything, one concludes that the change ΔL has a univariate normal distribution with variance $\mathbf{b}'\Sigma\mathbf{b}$, i.e. $\Delta L \sim \mathcal{N}(0, \mathbf{b}'\Sigma\mathbf{b})$. Here it is assumed that the dependence structure of the underwriting risk factor changes are governed by the covariance matrix Σ. Note that Σ can be decomposed into its correlation matrix R and a diagonal matrix Δ comprising the risk factor changes' standard deviations on the diagonal. Hence, $\Sigma = \Delta R \Delta$. Both the correlation coefficients and the standard deviations are based on expert opinion; no historical time series exists from which estimates could be inferred. Knowing the distribution of ΔL, one can apply a risk measure ϱ to arrive at a capital charge for the parametric insurance risk. Within the Swiss Solvency Test, one uses expected shortfall at the confidence level $\alpha = 0.99$.

Table 2 shows the correlation matrix R currently being used in the Swiss Solvency Test, whereas Table 3 contains the standard deviations of the underwriting risk factor changes.

Table 2 Correlation matrix R of the life insurance risk factor changes.

		Individual						Group					
		R1	R2	R3	R4	R5	R6	R1	R2	R3	R4	R5	R6
Individual	R1	1	0	0	0	0	0	1	0	0	0	0	0
	R2	0	1	0	0	0	0	0	1	0	0	0	0
	R3	0	0	1	0	0	0	0	0	1	0	0	0
	R4	0	0	0	1	0	0	0	0	0	1	0	0
	R5	0	0	0	0	1	0.75	0	0	0	0	1	0.75
	R6	0	0	0	0	0.75	1	0	0	0	0	0.75	1
Group	R1	1	0	0	0	0	0	1	0	0	0	0	0
	R2	0	1	0	0	0	0	0	1	0	0	0	0
	R3	0	0	1	0	0	0	0	0	1	0	0	0
	R4	0	0	0	1	0	0	0	0	0	1	0	0
	R5	0	0	0	0	1	0.75	0	0	0	0	1	0.75
	R6	0	0	0	0	0.75	1	0	0	0	0	0.75	1

Table 3 Standard deviations of the life insurance risk factor changes (in percentage).

	Individual							Group						
	R1	R2	R3	R4	R5	R6	R7	R1	R2	R3	R4	R5	R6	R7
σ_i	5	10	10	10	25	0	10	5	10	20	10	25	0	10

6.3 Non-life insurance risk

For the purpose of this article, the risk categories (in their general form) already discussed for life above, also apply. Clearly there are many distinctions at the product level. For instance, in non-life we often have contracts over a shorter time period, frequency risk may play a bigger role (think for instance of hail storms) and especially in the realm of catastrophe risk, numerous specific methods have been developed by non-life actuaries. Often techniques borrowed from (non-life) risk theory are taken over by the banking world. Examples are the modeling of loss distributions, the axiomatization of risk measures, IBNR and related techniques, Panjer recursion, etc. McNeil et al. (2005) yield an exhaustive overview on the latter techniques and refer to them as Insurance Analytics. For a comprehensive summary of the modeling of loss distributions, see for instance Klugman et al. (2004). An example stressing the interplay between financial and insurance risk is Schmock (1999).

In the Swiss Solvency Test non-life model the aim is to determine the change in risk bearing capital within one year due to the variability of the technical result. The model is based on the accident year principle. That is, claims are grouped according to the date of occurrence (and not accord-

ing to the date or year when they are reported). Denoting by $[T_0, T_1]$ with $T_1 = T_0 + 1$ the one-year time interval under consideration, the technical result within $[T_0, T_1]$ is not only determined by the claims occurring in this period, but also by the claims that have previously occurred and whose settlement stretches across $[T_0, T_1]$.
The current year claims are further grouped into high frequency-small severity claims ("small claims") and low frequency-high severity claims ("large claims"). It is stipulated that the total of small claims has a gamma distribution, whereas in the large claims regime a compound Poisson distribution with Pareto distributed claim sizes is used.

As for the claims that have occurred in the past and are not yet settled, the focus is on the annual reserving result; it is defined as the difference between the sum of the claim payments during $[T_0, T_1]$ plus the remaining provisions after T_1 minus the provisions that were originally set up at time T_0. Within the Swiss Solvency Test, this one-year reserve risk is modeled by means of a (shifted) log-normally distributed random variable.

To obtain the ultimate probability distribution of the non-life risk, one first aggregates the small claims and the large claims risk, thereby assuming independence between these two risk categories. A second convolution is then required to combine the resulting current year risk with the reserve risk, again assuming independence.

7 Aggregation of Risks

A key issue for the economic capital modeling is the aggregation of risks. Economic capital models are too often based on the tacit assumption that risk can be diversified via aggregation. For VaR in the context of very heavy-tailed distributions, however, the idea of a capital relief due to pooling of risks may shipwreck, see Example 1 on page 734 where it is shown that VaR is not sub-additive for an infinite mean model. The (non-) existence of subadditivity is closely related to Kolmogorov's strong law of large numbers, see Nešlehová et al. (2006).
In a formal way, diversification could be defined as follows:

Definition 3 Let $X_1, \ldots, X_n$ be a sequence of risks and ϱ a risk measure. Diversification is then expressed as

$$\mathcal{D}_\varrho := \sum_{k=1}^{n} \varrho(X_k) - \varrho\big(\sum_{k=1}^{n} X_k\big).$$

Extreme heavy-tailedness is one reason why VaR fails to be sub-additive. Another reason overthrowing the idea of diversification is extreme skewness of risks as the following simple example demonstrates. Assume that a loss of EUR 10 million or more is incurred with a probability of 3% and that the

loss will be EUR 100'000 with a probability of 97%. In this case the VaR at the 95% level is EUR 100'000, while aggregating two such independent losses yields a VaR of more than EUR 10 million.

The modeling of dependence is a central element in quantitative risk management. In most cases, the assumption of *independent* (market-) risky instruments governing the portfolio value is too simplistic and unrealistic. Correlation is by far the most used technique in modern finance and insurance to describe dependence between risks. And yet correlation is only one particular measure of stochastic dependence among others. Whereas correlation is perfectly suited for elliptically distributed risks, dangers lurk if correlation is used in a non-elliptical world. Recall that independence of two random variables always implies their uncorrelatedness. The converse, however, does in general not hold.

We have shown in Example 2 on page 734 that VaR is sub-additive in a normal risks regime. Indeed, this fact can be used to aggregate market and insurance risk in a variance-covariance type model, see Sections 4.1.1 and 6.2. There, the required economic capital when combining market and insurance risks will naturally be reduced compared to the stand-alone capital requirements.

The above example with extremely skewed risks also shows that independence can be worse than comonotonicity. Comonotonicity means that the risks $X_1, \ldots, X_d$ are expressible as increasing functions of a single random variable, Z say. In the case of comonotonic risks VaR is additive, see for instance McNeil et al. (2005), Proposition 6.15. For given marginal distribution functions and unknown dependence structure, it is in fact possible to calculate upper and lower bounds for VaR, see Embrechts et al. (2003). However, these bounds often prove inappropriate in many practical risk management applications. As a consequence, the dependence structure among the risks needs to be modeled explicitly – if necessary by making the appropriate assumptions.

8 Summary

In this paper, we have summarized some of the issues underlying the quantitative modeling of risks in insurance and finance. The taxonomy of risk discussed is of course incomplete and very much driven by the current supervisory process within the financial and insurance services industry. We have hardly vouched upon the huge world of risk mitigation via financial derivatives and alternative risk transfer, like for instance catastrophe bonds. Nor did we discuss in any detail specific risk classes like liquidity risk and model risk; for the latter, Gibson (2000) yields an introduction. Beyond the discussion of quantitative risk measurement and management, there is also an increasing awareness that qualitative aspects of risk need to be taken seri-

ously. Especially through the recent discussions around operational risk, this qualitative aspect of risk management became more important. Though modern financial and actuarial techniques have highly influenced the quantitative modeling of risk, there is also a growing awareness that there is an end to the line for this quantitative approach. Though measures like VaR and the whole statistical technology behind it have no doubt had a considerable influence on the handling of modern financial instruments, hardly anybody might believe that a single number like VaR can really summarize the overall complexity of risk in an adequate way. For operational risk, this issue is discussed in Nešlehová et al. (2006); see also Klüppelberg and Rootzén (1999).

Modern risk management is being applied to areas of industry well beyond the financial ones. Examples include the energy sector and the environment. Geman (2005) gives an overview of some of the modeling and risk management issues for these markets. A more futuristic view on the types of risk modern society may want to manage is given in Shiller (2003).

Acknowledgement The authors would like to thank Thomas Mikosch for a careful reading of a first version of the paper.

References

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999): Coherent measures of risk. *Mathematical Finance* **9**, 203–228.

Basel Committee on Banking Supervision (2005): *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Bank for International Settlements (BIS), Basel. `www.bis.org/publ/bcbs118.pdf`

Brummelhuis, R. and Guégan, D. (2000): Extreme values of conditional distributions of GARCH(1,1) processes. *Preprint, University of Reims*.

Brummelhuis, R. and Guégan, D. (2005): Multi period conditional distribution functions for conditionallly normal GARCH(1,1) models. *Journal of Applied Probability* **42**, 426–445.

Chavez-Demoulin, V., Embrechts, P. and Nešlehová, J. (2006): Quantitative models for operational risk: extremes, dependence and aggregation. *Journal of Banking and Finance* to appear.

Duffie, D. and Singleton, K.J. (1999): Modeling term structures of defaultable bonds. *Review of Financial Studies* **12**, 687–720.

Duffie, D. and Singleton, K.J. (2003): *Credit Risk. Pricing, Measurement and Management*. Princeton University Press.

Embrechts, P., Höing, A. and Juri, A. (2003): Using copulae to bound the Value-at-Risk for functions of dependent risks. *Finance and Stochastics* **7**, 145–167.

Geman, H. (2005): *Commodities and Commodity Derivatives: Modelling and Pricing for Agriculturals, Metals and Energy*. John Wiley, Chichester.

Gibson, R. (2000): *In: Gibson, R. (Ed.): Model Risk, Concepts, Calibration and Pricing*. Risk Waters Group, London.

Jorion, P. (2000): *Value at Risk*. McGraw-Hill, New York.

Kaufmann, R. (2005): Long-term risk management. *Proceedings of the 15th International AFIR Colloquium Zürich*.

Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2004): *Loss Models: From Data to Decisions* (2nd ed.). John Wiley, New York.

Klüppelberg, C. and Rootzén, H. (1999): A single number can't hedge against economic catastrophes. *Ambio* **28**, 550–555.

Lando, D. (1998): Cox processes and credit-risky securities. *Review of Derivatives Research* **2**, 99–120.

McNeil, A.J., Frey, R. and Embrechts, P. (2005): *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

Merton, R. (1974): On the pricing for corporate debt: the risk structure of interest rates. *Journal of Finance* **29**, 449–470.

Moscadelli, M. (2004): The modelling of operational risk: experience with the analysis of the data collected by the Basel committee. *Technical Report* **517**, Banca d'Italia.

Müller, H. (1997): *Solvency of insurance undertakings*. Conference of the Insurance Supervisory Authorities of the Member States of the European Union. `www.ceiops.org`

Nešlehová, J., Embrechts, P. and Chavez-Demoulin, V. (2006): Infinite-mean models and the LDA for operational risk. *Journal of Operational Risk* **1**, 3–25.

Sandström, A. (2005): *Solvency: Models, Assessment and Regulation*. Chapman & Hall/CRC, Boca Raton.

Schmock, U. (1999): Estimating the value of the Wincat coupons of the Winterthur insurance convertible bond: a study of the model risk. *ASTIN Bulletin* **29**, 101–163.

Sharma, P. (2002): Prudential supervision of insurance undertakings. *Conference of Insurance Supervisory Services of the Member States of the European Union*. `www.ceiops.org`

Shiller, R. J. (2003): *The New Financial Order: Risk in the Twenty-First Century*. Princeton University Press.

Value–at–Risk Models

Peter Christoffersen*

Abstract In this chapter, we build first a univariate and then a multivariate filtered historical simulation (FHS) model for financial risk management. Both the univariate and multivariate methods simulate future returns from a model using historical return innovations. While the former relies on portfolio returns filtered by a dynamic variance model, the latter uses individual or base asset return innovations from dynamic variance and correlation models. The univariate model is suitable for passive risk management or risk measurement whereas the multivariate model is useful for active risk management such as optimal portfolio allocation. Both models are constructed in such a way as to capture the stylized facts in daily asset returns and to be simple to estimate. The FHS approach enables the risk manager to easily compute Value-at-Risk and other risk measures including Expected Shortfall for various investment horizons that are conditional on current market conditions. The chapter also lists various alternatives to the suggested FHS approach.

1 Introduction and Stylized Facts

In this chapter, we apply some of the tools from previous chapters to develop a tractable dynamic model for computing the Value-at-Risk (VaR) and other risk measures of a portfolio of traded assets.

The VaR is defined as a number such that there is a probability p of exhibiting a worse return over the next K days, and where p and K must

Peter Christoffersen
McGill University, Desautels Faculty of Management, 1001 Sherbrooke Street West, Montreal, Quebec, Canada, H3A 1G5, e-mail: peter.christoffersen@mcgill.ca

* I am also affiliated with CIRANO, CIREQ, CREATES and ERM and I would like to thank FQRSC, IFM^2 and SSHRC for financial support.

DOI: 10.1007/978-3-540-71297-8_33,

be predetermined by the risk manager. The VaR is thus simply a quantile of the return distribution. Clearly, the quantile of a distribution does not tell us everything about risk. Importantly, it does not tell us how large the likely magnitude of losses is on those days when the return is worse than the VaR. Expected Shortfall (ES), which is defined as the expected return conditional on the return being worse than the VaR, has been suggested as an alternative to VaR and will also be discussed in this chapter. But, VaR remains by far the most common risk metric used in practice.

The so-called historical simulation (HistSim) method has emerged as the industry standard for computing VaR. It computes VaR in two simple steps. First, a series of hypothetical historical portfolio returns are constructed, using today's portfolio weights and historical asset returns. Second, the quantile of the hypothetical historical portfolio returns is computed. Advocates of the HistSim approach highlight its "model-free" nature. However it is clearly not "assumption-free". HistSim essentially assumes that asset returns are i.i.d. which is unfortunately not the case empirically.

The objective in this chapter is therefore to design a dynamic alternative to the static HistSim approach. Specifically, we wish to build a risk model with the following characteristics:

- The model is a fully specified data-generating process which can be estimated on daily returns
- The model can be estimated and implemented for portfolios with a large number of assets
- VaR can be easily computed for any prespecified level of confidence, p, and for any horizon of interest, K
- VaR is conditional on the current market conditions
- Risk measures other than the VaR can be calculated easily

To deliver accurate risk predictions, the model should reflect the following stylized facts of daily asset returns

- Daily returns have little or no exploitable conditional mean predictability
- The variance of daily returns greatly exceeds the mean
- The variance of daily returns is predictable
- Daily returns are not normally distributed
- Even after standardizing daily returns by a dynamic variance model, the standardized daily returns are not normally distributed
- Positive and negative returns of the same magnitude may have different impact on the variance
- Correlations between assets appear to be time-varying
- As the investment horizon increases, the return data distribution approaches the normal distribution

Again, the objective is to build a dynamic market risk management model that captures these salient features of daily asset returns, that contains only

few parameters to be estimated, and that is easily implemented on a large set of assets.

In Section 2, we will consider a univariate approach and, in Section 3, a multivariate approach to dynamic risk modeling. The univariate model simulates historical portfolio return shocks from a dynamic variance model, and the multivariate model simulates historical asset return shocks by means of both dynamic variance and correlation models. The univariate model is suitable for passive risk management or risk measurement, whereas the multivariate model is useful for active risk management such as optimal portfolio allocation. The end of each section will discuss alternatives to the approach taken here. Section 4 concludes.

2 A Univariate Portfolio Risk Model

In this section, we will consider a simple univariate approach to modeling the dynamic risk of a portfolio. Just as in the HistSim approach mentioned above, we consider a time series of T hypothetical historical portfolio returns computed using today's portfolio weights, and historical returns on n assets

$$\{r_t\}_{t=1}^{T} \equiv \left\{ \sum_{j=1}^{n} w_{T,j} r_{t,j} \right\}_{t=1}^{T},$$

where $r_{t,j}$ denotes the log return on asset j from the market close on day $t-1$ to market close on day t, that is, $r_{t,j} = \ln(S_{t,j}/S_{t-1,j})$, and where $w_{T,j}$ denotes today's weight of asset j in the portfolio.

The univariate risk model proceeds by simply modeling the properties of the univariate portfolio return, r_t. One great advantage of this approach is that the correlations and other interdependencies between the n assets do not need to be modeled. The downside however of the approach is that it is conditional on the portfolio weights. When these weights change, then so should the estimated risk model. This portfolio level approach is sometimes referred to as a *passive* risk model as it does not directly allow for studying the effects of actively managing the risk of the portfolio by changing the portfolio weights.

We proceed by making some key assumptions on the daily portfolio return process. We will assume that

$$r_t = \sigma_t z_t \qquad z_t \overset{i.i.d}{\sim} G(0,1), \tag{1}$$

where the dynamic return volatility σ_t is known at the end of day $t-1$, and where the independent and identically distributed (i.i.d.) return shock, z_t, is from a potentially nonnormal distribution, $G(0,1)$, with zero mean and unit

variance. Note that we set the return mean to zero, allowed for time-varying volatility, and for conditional nonnormality. These assumptions are all in line with the stylized facts outlined in Section 1. Note also, however, that we have ruled out time-varying conditional skewness and kurtosis which is sometimes found to be relevant in asset return modeling. See for example Hansen (1994) and Harvey and Siddique (1999).

We proceed by first modeling and estimating σ_t, and then, subsequently, moving on to the specification of $G(0,1)$.

2.1 The dynamic conditional variance model

In order to capture the time-varying volatility found in the daily returns, we rely on the NGARCH(1,1) model. In this model, the variance for date t can be computed based on the return and the variance for date $t-1$ as follows:

$$\sigma_t^2 = \omega + \alpha \left(r_{t-1} - \theta\sigma_{t-1}\right)^2 + \beta\sigma_{t-1}^2,$$

where a positive θ captures the fact that a negative return will increase variance by more than a positive return of the same magnitude. This asymmetry effect is one of the stylized facts listed in Section 1.

The unconditional –or long-run– variance in this model can be derived as

$$\sigma^2 = E\left[\sigma_t^2\right] = \frac{\omega}{1 - \alpha\left(1+\theta^2\right) - \beta} \equiv \frac{\omega}{\kappa},$$

where $\kappa \equiv 1 - \alpha\left(1+\theta^2\right) - \beta$ is interpretable as the speed of mean reversion in variance.

Setting $\omega = \sigma^2\kappa$ and substituting it into the dynamic variance equation yields

$$\begin{aligned}\sigma_t^2 &= \sigma^2\kappa + \alpha\left(r_{t-1} - \theta\sigma_{t-1}\right)^2 + \beta\sigma_{t-1}^2 \\ &= \sigma_{t-1}^2 + \kappa\left(\sigma^2 - \sigma_{t-1}^2\right) + \alpha\left(r_{t-1}^2 - \sigma_{t-1}^2 - 2\theta r_{t-1}\sigma_{t-1}\right),\end{aligned} \tag{2}$$

where, in the second line, we have simply expanded the square and applied the definition of κ.

The advantage of writing the model in this form is two-fold. First, we can easily impose the long-run variance, σ^2, to be the sample variance, before estimating the other parameters. This is referred to as variance targeting. Second, we can easily impose variance stationarity on the model, by ensuring that $\kappa > 0$ when estimating the remaining parameters. Finally, we guarantee variance positivity by forcing $\alpha > 0$ when estimating the parameters.

The parameters $\{\kappa, \alpha, \theta\}$ that determine the volatility dynamics are easily estimated by numerically optimizing the quasi maximum likelihood criterion of the estimation sample

$$QMLE\left(\kappa,\alpha,\theta\right)=-\frac{1}{2}\sum_{t=1}^{T}\left(\ln\left(\sigma_t^2\right)+r_t^2/\sigma_t^2\right). \tag{3}$$

Typically, κ is found to be close to zero reflecting slow mean reversion and thus high predictability in the daily variance. The autocorrelation function of the absolute shock, $|z_t|=|r_t/\sigma_t|$, provides a useful diagnostic of the volatility model.

2.2 Univariate filtered historical simulation

We now turn to the specification of the distribution $G(0,1)$ of the return shock, z_t. The easiest way to proceed would be to assume that the shocks follow the standard normal distribution. As the standard normal distribution has no parameters, the specification of the model would then be complete and the model ready for risk forecasting. From the list of stylized facts in Section 1, we know however that the assumption of a normal distribution is not appropriate for most speculative assets at the daily frequency.

The question which alternative distribution to choose then arises? Rather than forcing such a choice, we here rely on a simple resampling scheme, which, in financial risk management, is sometimes referred to as filtered historical simulation (FHS). The term "filtered" refers to the fact that we are not simulating from the set of raw returns, but from the set of shocks, z_t, which are returns filtered by the GARCH model.

It is simple to construct a one-day VaR from FHS. We calculate the percentile of the set of historical shocks, $\{z_t\}_{t=1}^{T}$, where $z_t=r_t/\sigma_t$, and multiply that onto the one-day ahead volatility

$$VaR_{T,1}^{p}=\sigma_{T+1}Percentile\left\{\{z_t\}_{t=1}^{T},100p\right\}. \tag{4}$$

where the $Percentile$ function returns a number, z_p, such that $100p$ percent of the numbers in the set $\{z_t\}_{t=1}^{T}$ are smaller than z_p. Note that, by construction of the GARCH model, the one-day-ahead volatility is known at the end of the previous day, so that $\sigma_{T+1|T}=\sigma_{T+1}$ and we simply use the latter simpler notation. The Expected Shortfall for the one-day horizon can be calculated as

$$ES_{T,1}^{p}=\sigma_{T+1}\frac{1}{p*T}*\sum_{t=1}^{T}z_t*\mathbf{1}\left(z_t<VaR_{T,1}^{p}/\sigma_{T+1}\right),$$

where $\mathbf{1}(*)$ denotes the indicator function returning a 1 if the argument is true, and zero otherwise.

When computing a multi-day ahead VaR, the GARCH variance process must be simulated forward using random draws, $z_{i,k}$, from the historical

shocks, $\{z_t\}_{t=1}^{T}$. The random drawing can be operationalized by generating a discrete uniform random variable which is distributed from 1 to T. Each draw from the discrete distribution then tells us which shock to select. We build up a distribution of hypothetical future returns as

$$
\begin{array}{ccccccc}
 & & z_{1,1} & \rightarrow & r_{1,T+1} & \rightarrow & \sigma^2_{1,T+2} \quad \cdots \\
 & \nearrow & & & \cdots & & \cdots \\
\sigma^2_{T+1} & \longrightarrow & z_{i,1} & \rightarrow & r_{i,T+1} & \rightarrow & \sigma^2_{i,T+2} \quad \cdots \\
 & \searrow & & & \cdots & & \cdots \\
 & & z_{M,1} & \rightarrow & r_{M,T+1} & \rightarrow & \sigma^2_{M,T+2} \quad \cdots
\end{array}
$$

$$
\begin{array}{ccccccccc}
z_{1,k} & \rightarrow & r_{1,T+k} & \rightarrow & \sigma^2_{1,T+k+1} & \cdots & z_{1,K} & \rightarrow & r_{1,T+K} \\
 & & \cdots & & & \cdots & & & \cdots \\
z_{i,k} & \rightarrow & r_{i,T+k} & \rightarrow & \sigma^2_{i,T+k+1} & \cdots & z_{i,K} & \rightarrow & r_{i,T+K} \\
 & & \cdots & & & \cdots & & & \cdots \\
z_{M,k} & \rightarrow & r_{M,T+k} & \rightarrow & \sigma^2_{M,T+k+1} & \cdots & z_{M,K} & \rightarrow & r_{M,T+K}
\end{array}
$$

where $r_{i,T+k}$ is the return for day $T+k$ on simulation path i, M is the number of times we draw with replacement from the T standardized returns on each future date, and K is the horizon of interest. At each time step, the GARCH model in (2) is used to update the conditional variance and the return model in (1) is used to construct returns from shocks.

We end up with M sequences of hypothetical daily returns for day $T+1$ through day $T+K$. From these hypothetical daily returns, we calculate the hypothetical $K-$day returns as

$$r_{i,T:K} = \sum_{k=1}^{K} r_{i,T+k}, \text{ for } i = 1, 2, ..., M.$$

If we collect the M hypothetical K-day returns in a set $\{r_{i,T:K}\}_{i=1}^{M}$, then we can calculate the $K-$day Value at Risk simply by calculating the $100p$ percentile as in

$$VaR^{p}_{T,K} = Percentile\left\{\{r_{i,T:K}\}_{i=1}^{M}, 100p\right\}.$$

At this point it is natural to ask how many simulations, M, are needed? Ideally, M should of course be as large as possible in order to approximate closely the true but unknown distribution of returns and thus the VaR. On modern computers, taking $M = 100,000$ is usually not a problem and would yield on average $1,000$ tail observations when computing a VaR for $p = 0.01$. It is important to note that the smaller the p the larger an M is needed in order to get a sufficient number of extreme tail observations.

The ES measure can be calculated from the simulated returns by taking the average of all the $r_{i,T:K}$ that fall below the $VaR^p_{T,K}$ number, that is

$$ES^p_{T,K} = \frac{1}{p*M} * \sum_{i=1}^{M} r_{i,T:K} * \mathbf{1}\left(r_{i,T:K} < VaR^p_{T,K}\right).$$

The advantages of the FHS approach are threefold. First, it captures current market conditions by means of the volatility dynamics. Second, no assumptions need to be made on the distribution of the return shocks. Third, the method allows for the computation of any risk measure for any investment horizon of interest.

2.3 Univariate extensions and alternatives

The GARCH model that we used in (2) is taken from Engle and Ng (1993). Andersen, Bollerslev, Christoffersen and Diebold (2006a) survey the range of viable volatility forecasting approaches. The filtered historical simulation approach in (4) was suggested by Barone-Adesi, Bourgoin, and Giannopoulos (1998), Diebold, Schuermann, and Stroughair (1998), and Hull and White (1998).

The general univariate model in (1) and (2) contains a number of standard risk models as special cases:

- The i.i.d. Normal model where $G(0,1) = N(0,1)$ and $\kappa = \alpha = \theta = 0$
- The RiskMetrics model where $G(0,1) = N(0,1)$ and $\kappa = 0$ and $\theta = 0$
- The GARCH-Normal where $G(0,1) = N(0,1)$
- The GARCH-CF where $G^{-1}(0,1)$ is approximated using the Cornish-Fisher approach
- The GARCH-EVT model where the tail of $G(0,1)$ is specified using extreme value theory
- The GARCH-t(d) where $G(0,1)$ is a standardized Student's t distribution

As discussed in Christoffersen (2003), these models can be estimated relatively easily using a variant of the likelihood function in (3) or by matching moments of z_t with model moments. However, they all contain certain drawbacks that either violate one or more of the stylized facts listed in Section 1, or that fail to meet one or more of the objectives listed in Section 1 as well: The i.i.d. Normal model does not allow for variance dynamics. The RiskMetrics model (JP Morgan, 1996) does not aggregate over time to normality nor does it capture the leverage effect. The GARCH-Normal does not allow for conditional nonnormality, and the GARCH-CF and GARCH-EVT (McNeill and Frey, 2000) models are not fully specified data-generating processes. The GARCH-t(d) (Bollerslev, 1987) comes closest to meeting our objectives but

needs to be modified to allow for conditional skewness. See, for example, Hansen (1994).

Some quite different approaches to VaR estimation have been suggested. The Weighted Historical Simulation approach in Bodoukh, Richardson and Whitelaw (1998) puts higher probability on recent observations when computing the HistSim VaR. However, see Pritsker (2001) for a critique. The CaViaR approach in Engle and Manganelli (2004) and the dynamic quantile approach in Gourieroux and Jasiak (2006) model the return quantile directly rather than specifying a complete data generating process. Finally, note that Manganelli (2004) suggests certain univariate models for approximate portfolio allocation by variance sensitivity analysis.

3 Multivariate, Base–Asset Return Methods

The univariate methods discussed in Section 2 are useful if the main purpose of the risk model is risk measurement. If instead the model is required for active risk management including deciding on optimal portfolio allocations, or VaR sensitivities to allocation changes, then a multivariate model may be required. In this section, we build on the model in Section 2 to develop a fully specified large-scale multivariate risk model.

We will assume that the risk manager knows his set of assets of interest. This set can either contain all the assets in the portfolio or a smaller set of so-called base assets which are believed to be the main drivers of risk in the portfolio. Base asset choices are, of course, portfolio-specific, but typical examples include equity indices, bond indices, and exchange rates as well as more fundamental economic drivers such as oil prices and real estate prices. Regression analysis can be used to assess the relationship between each individual asset and the base assets.

Once the set of assets has been determined, the next step in the multivariate model is to estimate a dynamic volatility model of the type in Section 1 for each of the n assets. When this is complete, we can write the n base asset returns in vector form

$$R_t = D_t Z_t,$$

where D_t is an n by n diagonal matrix containing the GARCH standard deviations on the diagonal, and zeros on the off diagonal. The n by 1 vector Z_t contains the shocks from the GARCH models for each asset.

Now, define the conditional covariance matrix of the returns as

$$Var_{t-1}(R_t) = \Sigma_t = D_t \Gamma_t D_t,$$

where Γ_t is an n by n matrix containing the base asset correlations on the off diagonals and ones on the diagonal. The next step in the multivariate model is to develop a tractable model for Γ_t.

3.1 The dynamic conditional correlation model

We wish to capture time variation in the correlation matrix of base asset returns without having to estimate many parameters. The correlation matrix has $n(n-1)/2$ unique elements but the dynamic conditional (DCC) model offers a convenient framework for modeling these using only two parameters that require numerical estimation methods.

The correlation dynamics are modeled through past cross products of the shocks in Z_t

$$\begin{aligned} Q_t &= \Omega + \alpha \left(Z_{t-1} Z_{t-1}' \right) + \beta Q_{t-1} \\ &= Q \left(1 - \alpha - \beta \right) + \alpha \left(Z_{t-1} Z_{t-1}' \right) + \beta Q_{t-1} \\ &= Q_{t-1} + \kappa \left(Q - Q_{t-1} \right) + \alpha \left(Z_{t-1} Z_{t-1}' - Q_{t-1} \right), \end{aligned} \tag{5}$$

where we have used

$$E\left[Q_t \right] \equiv Q = \Omega / \left(1 - \alpha - \beta \right) \equiv \Omega / \kappa.$$

The unconditional sample covariance matrix of Z_t provides an estimate of Q, leaving only κ and α to be estimated by numerical optimization. Forcing $\kappa > 0$ in estimation ensures correlation stationarity.

The conditional correlations in Γ_t are given by standardizing the relevant elements of the Q_t matrices. Let $\rho_{ij,t}$ be the correlation between asset i and asset j on day t. Then we have

$$\rho_{ij,t} = \frac{q_{ij,t}}{\sqrt{q_{ii,t} q_{jj,t}}}, \tag{6}$$

where $q_{ij,t}$, $q_{ii,t}$, and $q_{jj,t}$ are elements of Q_t.

The dynamic correlation parameters κ and α can now be estimated by maximizing the QMLE criterion on the multivariate sample

$$QMLE\left(\kappa, \alpha \right) = -\frac{1}{2} \sum_{t=1}^{T} \left(\log \left(\| \Gamma_t \| \right) + Z_t' \Gamma_t^{-1} Z_t \right),$$

where $\| \Gamma_t \|$ denotes the determinant of Γ_t.

3.2 Multivariate filtered historical simulation

Based on the stylized facts in Section 1, we do not want to assume that the shocks to the assets are normally distributed. Nor do we wish to assume that they stem from the same distribution. Instead, we will simulate from

historical shocks asset by asset to compute forward-looking VaRs and other risk measures.

We first create a database of historical dynamically uncorrelated shocks from which we can resample. We create the dynamically uncorrelated historical shock as

$$Z_t^D = \Gamma_t^{-1/2} Z_t,$$

where, $\Gamma_t^{-1/2}$ is the inverse of the matrix square-root of the conditional correlation matrix Γ_t. The matrix square root, $\Gamma_t^{1/2}$, can be computing using the spectral decomposition of Γ_t. In their chapter in this Handbook, Patton and Sheppard (2008) recommend the spectral decomposition over the standard Cholesky decomposition because the latter is not invariant to the ordering of the return variables in the vector Z_t.

When calculating the multi-day conditional VaR and other risk measures from the model, we need to simulate daily returns forward from today's (day T's) forecast of tomorrow's matrix of volatilities, D_{T+1} and correlations, Γ_{T+1}. The returns are computed from the GARCH and DCC models above.

From the data base of uncorrelated shocks $\left\{Z_t^D\right\}_{t=1}^T$, we can draw a random vector of historical uncorrelated shocks, called $Z_{i,T+1}^D$. It is important to note that in order to preserve asset-specific characteristics and potential extreme correlation in the shocks, we draw an entire vector representing the same day for all the assets.

From this draw, we can compute a random return for day $T+1$ as

$$\begin{aligned} R_{i,T+1} &= D_{T+1} \Gamma_{T+1}^{1/2} Z_{i,T+1}^D \\ &= D_{T+1} Z_{i,T+1}. \end{aligned}$$

Using the simulated shock vector, $Z_{i,T+1}$, we can now update the volatilities and correlations using the GARCH model in (2) and the DCC model in (5) and (6). We thus obtain $D_{i,T+2}$ and $\Gamma_{i,T+2}$. Drawing a new vector of uncorrelated shocks, $Z_{i,T+2}^D$, enables us to simulate the return for the second day as

$$\begin{aligned} R_{i,T+2} &= D_{i,T+2} \Gamma_{i,T+2}^{1/2} Z_{i,T+2}^D \\ &= D_{i,T+2} Z_{i,T+2}. \end{aligned}$$

We continue this simulation for K days, and repeat it for $i = 1, ..., M$ simulated shocks.

The cumulative K-day log returns are calculated as

$$R_{i,T:K} = \sum_{k=1}^{K} R_{i,T+k}.$$

The portfolio Value-at-Risk (VaR) is calculated by computing the user-specified percentile of the M simulated returns for each horizon as in

$$VaR^{p}_{T,K} = Percentile\left\{\{W'_T R_{i,T:K}\}_{i=1}^{M}, 100p\right\},$$

where W_T is the vector of portfolio weights at the end of day T.

The Expected Shortfall (ES) is computed by taking the average of those simulated returns which are worse than the VaR

$$ES^{p}_{T,K} = \frac{1}{p * M}\sum_{i=1}^{M} W'_T R_{i,T:K} * \mathbf{1}\left(W'_T R_{i,T:K} < VaR^{p}_{T,K}\right).$$

The advantages of the multivariate FHS approach tally with those of the univariate case: It captures current market conditions by means of dynamic variance and correlation models. It makes no assumption on the conditional multivariate shock distributions. And, it allows for the computation of any risk measure for any investment horizon of interest.

3.3 Multivariate extensions and alternatives

The DCC model in (5) is due to Engle (2002). See also Tse and Tsui (2002). Extensions to the basic model are developed in Capiello, Engle and Sheppard (2004). For alternative multivariate GARCH approaches, see the surveys in Andersen, Bollerslev, Christoffersen and Diebold (2006a and b), and Bauwens, Laurent, and Rombouts (2006). Jorion (2006) discusses the choice of base assets.

Parametric alternatives to the filtered historical simulation approach include specifying a multivariate normal or Student's t distribution for the GARCH shocks. See, for example Pesaran and Zaffaroni (2004). The multivariate normal and Student's t asset distributions offer the advantage that they are closed under linear transformations so that the portfolio returns will be normal and Student's t, respectively, as well.

The risk manager can also specify parametric conditional distributions for each asset and then link these marginal distributions together to form a multivariate distribution by using a copula function. See, for example, Demarta and McNeil (2005), Patton (2004, 2006), and Jondeau and Rockinger (2005). The results in Joe (1997) suggest that the DCC model itself can be viewed as a copula approach. Multivariate versions of the extreme value approach have also been developed. See, for example, Longin and Solnik (2001), and Poon, Rockinger, and Tawn (2004).

4 Summary and Further Issues

In this chapter, we have built first a univariate and then a multivariate filtered historical simulation model for financial risk management. The models are constructed to capture the stylized facts in daily asset returns, they are simple to estimate, and they enable the risk manager to easily compute Value-at-Risk and other risk measures including Expected Shortfall for various investment horizons conditional on current market conditions. The univariate model is suitable for passive risk management or risk measurement whereas the multivariate model is useful for active risk management such as optimal portfolio allocation. We also discuss various alternatives to the suggested approach.

Because our focus has been on the modeling of *market risk*, that is the risk from fluctuations in observed market prices, other important types of risk have been left unexplored.

We have focused on applications where a relatively long history of daily closing prices is available for each asset or base asset. In practice, portfolios often contain assets where daily historical market prices are not readily observable. Examples include derivatives, bonds, loans, new IPOs, private placements, hedge funds, and real estate investments. In these cases, asset pricing models are needed to link the unobserved asset prices to prices of other liquid assets. The use of pricing models to impute asset prices gives rise to an additional source of risk, namely *model risk*. Hull and Suo (2002) suggest a method to assess model risk.

In loan portfolios, nontraded *credit risk*is the main source of uncertainty. Lando (2004) provides tools for credit risk modeling and credit derivative valuation.

Illiquidity can itself be a source of risk. Historical closing prices may be available for many assets but if little or no trade was actually conducted at those prices then the historical information may not properly reflect risk. In this case, *liquidity risk* should be accounted for. See Persaud (2003) for various aspects of liquidity risk.

Even when computing the VaR and ES for readily observed assets, the use of parametric models implies estimation risk which we have not accounted for here. Christoffersen and Goncalves (2005) show that estimation risk can be substantial, and suggest ways to measure it in dynamic models.

References

Andersen, T.G., Bollerslev, T., Christoffersen, P. and Diebold, F.X. (2006a): Volatility and Correlation Forecasting. *In: Elliott, G., Granger, C. and Timmermann, A. (Eds.): Handbook of Economic Forecasting*. North-Holland, Amsterdam.

Andersen, T.G., Bollerslev, T., Christoffersen, P. and Diebold, F.X. (2006b): Practical Volatility and Correlation Modeling for Financial Market Risk Management. *In:*

Carey, M. and Stulz, R. (Eds.): The Risks of Financial Institutions. University of Chicago Press.

Barone-Adesi, G., Bourgoin, F. and Giannopoulos, K. (1998): Don't Look Back. *Risk* **11**, 100–104.

Bauwens, L., Laurent, S. and Rombouts, J. (2006): Multivariate GARCH Models: a Survey. *Journal of Applied Econometrics* **21**, 79–109.

Bodoukh, J., Richardson, M., and Whitelaw, R. (1998): The Best of Both Worlds. *Risk* **11**, 64–67.

Bollerslev, T. (1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T. (1987): A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics* **69**, 542–547.

Cappiello, L., Engle, R.F. and Sheppard, K. (2004): Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns. *Manuscript, Stern School of Business New York University*.

Christoffersen, P. (2003): *Elements of Financial Risk Management*. Academic Press, San Diego.

Christoffersen, P. and Diebold, F. (2000): How Relevant is Volatility Forecasting for Financial Risk Management? *Review of Economics and Statistics* **82**, 1–11.

Christoffersen, P. and Goncalves, S. (2005): Estimation Risk in Financial Risk Management. *Journal of Risk* **7**, 1–28.

Christoffersen, P., Diebold, F. and Schuermann, T. (1998): Horizon Problems and Extreme Events in Financial Risk Management. *Economic Policy Review* Federal Reserve Bank of New York, October, 109–118.

Demarta, S., and McNeil, A. J. (2005): The t Copula and Related Copulas. *International Statistical Review* **73**, 111–129.

Diebold, F.X., Schuermann, T. and Stroughair, J. (1998): Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management. *In: Refenes, A.-P. N., Burgess, A.N. and Moody, J.D. (Eds.): Decision Technologies for Computational Finance*, 3–12. Kluwer Academic Publishers, Amsterdam.

Duffie, D. and Pan, J. (1997): An Overview of Value at Risk. *Journal of Derivatives* **4**, 7–49.

Engle, R. (1982): Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation. *Econometrica* **50**, 987–1008.

Engle, R. (2002): Dynamic Conditional Correlation - A Simple Class of Multivariate GARCH Models. *Journal of Business and Economic Statistics* **20**, 339–350.

Engle, R. and Manganelli, S. (2004): CAViaR: Conditional Autoregressive Value at Risk by Quantile Regression. *Journal of Business and Economic Statistics* **22**, 367–381.

Engle, R. and Ng, V. (1993): Measuring and Testing the Impact of News on Volatility. *Journal of Finance* **48**, 1749–1778.

Engle, R. F. and Sheppard, K. (2001): Theoretical and Empirical properties of Dynamic Conditional Correlation Multivariate GARCH. *NBER Working Paper* **8554**.

Gourieroux, C. and Jasiak, J. (2006): Dynamic Quantile Models. *Manuscript, University of Toronto*.

Hansen, B. (1994): Autoregressive Conditional Density Estimation. *International Economic Review* **35**, 705–730.

Harvey, C.R. and Siddique, A. (1999): Autoregressive Conditional Skewness. *Journal of Financial and Quantitative Analysis* **34**, 465–488.

Hull, J. and Suo, W. (2002): A methodology for assessing model risk and its application to the implied volatility function model. *Journal of Financial and Quantitative Analysis* **37**, 297–318.

Hull, J. and White, A. (1998): Incorporating Volatility Updating into the Historical Simulation Method for VaR. *Journal of Risk* **1**, 5–19.

Joe, H. (1997): *Multivariate Models and Dependence Concepts*. Chapman Hall, London.

Jondeau, E. and Rockinger, M. (2005): The Copula-GARCH Model of Conditional Dependencies: An International Stock-Market Application. *Journal of International Money and Finance* forthcoming.

Jorion, P. (2006): Value-at-Risk: The New Benchmark for Managing. *Financial Risk*. McGraw Hill, New York.

Morgan, J.P. (1996): *RiskMetrics – Technical Document* 4th Edition. New York.

Lando, D. (2004): *Credit Risk Modeling: Theory and Applications* Princeton University Press, New Jersey.

Longin, F. and Solnik, B. (2001): Extreme Correlation of International Equity Markets. *Journal of Finance* **56**, 649–676.

Manganelli, S. (2004): Asset Allocation by Variance Sensitivity Analysis. *Journal of Financial Econometrics* **2**, 370–389.

McNeil, A. and Frey, R. (2000): Estimation of Tail-Related Risk Measures for Heteroskedastic Financial Time Series: An Extreme Value Approach. *Journal of Empirical Finance* **7**, 271–300.

Patton, A. (2004): On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation. *Journal of Financial Econometrics* **2**, 130–168.

Patton, A. (2006): Modeling Asymmetric Exchange Rate Dependence. *International Economic Review* **47**, 527–556.

Patton, A.J. and Sheppard, K. (2008): Evaluating volatility and Correlation forecasts. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 801–838. Springer Verlag, New York.

Persaud, A. (2003): *Liquidity Black Holes: Understanding, Quantifying and Managing Financial Liquidity Risk*. Risk Books, London.

Pesaran, H. and Zaffaroni, P. (2004): Model Averaging and Value-at-Risk based Evaluation of Large Multi Asset Volatility Models for Risk Management. *Manuscript, University of Cambridge*.

Poon, S.-H., Rockinger, M. and Tawn, J. (2004): Extreme Value Dependence in Financial Markets: Diagnostics, Models and Financial Implications. *Review of Financial Studies* **17**, 581–610.

Pritsker, M. (2001): The Hidden Dangers of Historical Simulation. *Finance and Economics Discussion Series* **2001-27**. *Washington: Board of Governors of the Federal Reserve System*.

Tse, Y.K. and Tsui, K.C. (2002): A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-varying Correlations. *Journal of Business and Economic Statistics* **20**, 351–362.

Copula–Based Models for Financial Time Series

Andrew J. Patton

Abstract This paper presents an overview of the literature on applications of copulas in the modelling of financial time series. Copulas have been used both in multivariate time series analysis, where they are used to characterize the (conditional) cross-sectional dependence between individual time series, and in univariate time series analysis, where they are used to characterize the dependence between a sequence of observations of a scalar time series process. The paper includes a broad, brief, review of the many applications of copulas in finance and economics.

1 Introduction

The central importance of risk in financial decision-making directly implies the importance of dependence in decisions involving more than one risky asset. For example, the variance of the return on a portfolio of risky assets depends on the variances of the individual assets and also on the linear correlation between the assets in the portfolio. More generally, the distribution of the return on a portfolio will depend on the univariate distributions of the individual assets in the portfolio and on the dependence between each of the assets, which is captured by a function called a 'copula'.

The number of papers on copula theory in finance and economics has grown enormously in recent years. One of the most influential of the 'early' papers on copulas in finance is that of Embrechts, McNeil and Straumann (2002), which was circulated as a working paper in 1999. Since then, scores of papers have been written, exploring the uses of copulas in finance, macroeconomics, and

Andrew J. Patton
Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom, e-mail: andrew.patton@economics.ox.ac.uk

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_34,

microeconomics, as well as developing the estimation and evaluation theory required for these applications. Nelsen (2006) and Joe (1997) provide detailed and readable introductions to copulas and their statistical and mathematical foundations, while Cherubini, *et al.* (2004) focus primarily on applications of copulas in mathematical finance and derivatives pricing. In this survey I focus on financial time series applications of copulas.

A copula is a function that links together univariate distribution functions to form a multivariate distribution function. If all of the variables are continuously distributed,[1] then their copula is simply a multivariate distribution function with *Uniform*$(0,1)$ univariate marginal distributions. Consider a vector random variable, $\mathbf{X} = [X_1, X_2, ..., X_n]'$, with joint distribution $\mathbf{F}$ and marginal distributions $F_1, F_2, ..., F_n$. Sklar's (1959) theorem provides the mapping from the individual distribution functions to the joint distribution function:

$$\mathbf{F}(\mathbf{x}) = \mathbf{C}(F_1(x_1), F_2(x_2), ..., F_n(x_n)), \quad \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{1}$$

From any multivariate distribution, $\mathbf{F}$, we can extract the marginal distributions, F_i, and the copula, $\mathbf{C}$. And, more useful for time series modelling, given any set of marginal distributions $(F_1, F_2, ..., F_n)$ and any copula $\mathbf{C}$, equation (1) can be used to obtain a joint distribution with the given marginal distributions. An important feature of this result is that the marginal distributions do not need to be in any way similar to each other, nor is the choice of copula constrained by the choice of marginal distributions. This flexibility makes copulas a potentially useful tool for building econometric models.

Since each marginal distribution, F_i, contains all of the univariate information on the individual variable X_i, while the joint distribution $\mathbf{F}$ contains all univariate *and* multivariate information, it is clear that the information contained in the copula $\mathbf{C}$ must be all of the dependence information between the X_i's[2]. It is for this reason that copulas are sometimes known as 'dependence functions', see Galambos (1978). Note that if we define U_i as the 'probability integral transform' of X_i, i.e. $U_i \equiv F_i(X_i)$, then $U_i \sim Uniform(0,1)$, see Fisher (1932), Casella and Berger (1990) and Diebold, *et al.* (1998). Further, it can be shown that $\mathbf{U} = [U_1, U_2, ..., U_n]' \sim \mathbf{C}$, the copula of $\mathbf{X}$.

[1] Almost all applications of copulas in finance and economics assume that that variables of interest are continuously distributed. Notable exceptions to this include Heinen and Rengifo (2003) and Grammig, *et al.* (2004). The main complication that arises when considering marginal distributions that are not continuous is that the copula is then only uniquely defined on the Cartesian product of supports of the marginal distributions. Obtaining a copula that is defined on $\mathbb{R}^n$ requires an interpolation method. See Denuit and Lambert (2005) for one such method.

[2] It is worth noting that some dependence measures of interest in finance, and elsewhere, depend on both the copula *and* the marginal distributions; standard linear correlation is the leading example. Depending on one's orientation, and the application at hand, this is either a drawback of such dependence measures or a drawback of copula theory.

If the joint distribution function is n-times differentiable, then taking the n^{th} cross-partial derivative of equation (1) we obtain:

$$\begin{aligned}\mathbf{f}(\mathbf{x}) &\equiv \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n}\mathbf{F}(\mathbf{x}) \\ &= \prod_{i=1}^{n} f_i(x_i) \cdot \frac{\partial^n}{\partial u_1 \partial u_2 \cdots \partial u_n}\mathbf{C}(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \\ &\equiv \prod_{i=1}^{n} f_i(x_i) \cdot \mathbf{c}(F_1(x_1), F_2(x_2), ..., F_n(x_n)), \end{aligned} \tag{2}$$

and so the joint density is equal to the product of the marginal densities and the 'copula density', denoted **c**. This of course also implies that the joint log-likelihood is simply the sum of univariate log-likelihoods and the 'copula log-likelihood', which is useful in the estimation of copula-based models:

$$\log \mathbf{f}(\mathbf{x}) = \sum_{i=1}^{n} \log f_i(x_i) + \log \mathbf{c}(F_1(x_1), F_2(x_2), ..., F_n(x_n)). \tag{3}$$

The decomposition of a joint distribution into its marginal distributions and copula allows the researcher a great deal of flexibility in specifying a model for the joint distribution. This is clearly an advantage when the shape and goodness-of-fit of the model for the joint distribution is of primary interest. In situations where the researcher has accumulated knowledge about the distributions of the *individual* variables and wants to use that in constructing a joint distribution, copulas also have a valuable role. In other situations, for example when the researcher is primarily focussed on the conditional mean and/or conditional variance of a vector of variables, copulas may not be the 'right tool for the job', and more standard vector autoregressive models and/or multivariate GARCH models, see Silvennoinen and Teräsvirta (2008), may be more appropriate. For a lively discussion of the value of copulas in statistical modelling of dependence, see Mikosch (2006) and the associated discussion (in particular that of Embrechts, Joe, and Genest and Rémillard) and rejoinder.

To illustrate the potential of copulas for modelling financial time series, I show in Figure 1 some bivariate densities constructed using Sklar's theorem. All have $F_1 = F_2 = N(0,1)$, while I vary **C** across different parametric copulas,[3] constraining the linear correlation to be 0.5 in all cases. The upper left plot shows the familiar elliptical contours of the bivariate Normal density

[3] The Normal and Student's t copulas are extracted from bivariate Normal and Student's t distributions. The Clayton and Gumbel copulas are discussed in Nelsen (2006), equations 4.2.1 and 4.2.4 respectively. The symmetrised Joe-Clayton (SJC) copula was introduced in Patton (2006a) and is parameterised by the upper and lower tail dependence coeffficients, τ^U and τ^L. The mixed Normal copula is an equally-weighted mixture of two Normal copulas with parameters ρ_1 and ρ_2 respectively.

(with Normal marginals and a Normal copula), while the other plots show some of the flexibility that various copula models can provide. To quantify

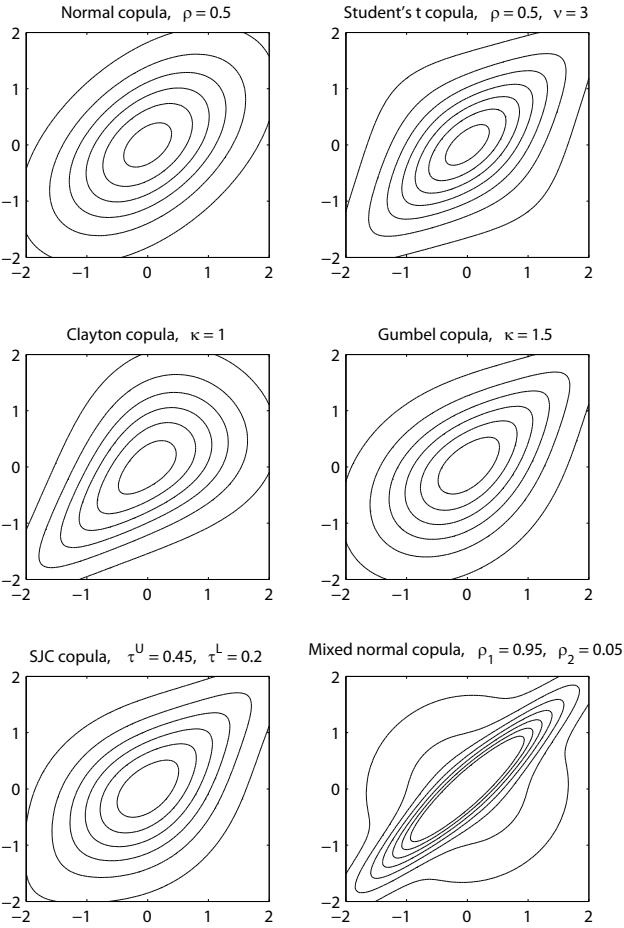

Fig. 1 Iso-probability contour plots for joint distributions with $N(0,1)$ marginal distributions and linear correlation coefficients of 0.5.

the different dependence structures provided by each copula, we can compare the dependence measures implied by each of these distributions, see Table 1. 'Quantile dependence' is related to a measure due to Coles, *et al.* (1999), and measures the probability of two variables both lying above or below a given quantile of their univariate distributions. It is defined as $\tau(q) = C(q,q)/q$ for $q \leq 1/2$ and $\tau(q) = \{1 - 2q + C(q,q)\}/(1-q)$ for $q > 1/2$. Lower and upper tail dependence can be defined as the limits of the quantile dependence

measures: $\lim_{q\to 0}\tau(q)=\tau^L$ and $\lim_{q\to 1}\tau(q)=\tau^U$, if the limits exist, which they do for the six copulas presented here.

Figure 1 and Table 1 show the variety of joint densities that may be constructed using copulas, even when we impose that both margins are Normal and that the correlation coefficient is 0.5. In many financial applications differences in, for example, lower tail dependence will have important implications. For example, if two assets have the Student's t copula rather than the Normal copula, then the probability of both asset returns lying below their lower 5% quantile (i.e., their 5% Value-at-Risk, see Embrechts, *et al.* 2008, and Christoffersen, 2008) is 0.37 rather than 0.24, meaning that a portfolio of these two assets will exhibit more extreme returns than identical assets with a Normal copula.

Table 1: Measures of dependence for joint distributions with various copulas

				Tail Dependence		5% Quantile Dependence	
			Linear				
Copula	Parameter(s)		Correlation	Upper	Lower	Upper	Lower
Normal(ρ)	0.5	–	0.50	0.00	0.00	0.24	0.24
Student's $t(\rho,\nu)$	0.5	3	0.50†	0.31	0.31	0.37†	0.37†
Clayton(κ)	1	–	0.50†	0.00	0.50	0.10	0.51
Gumbel(κ)	1.5	–	0.50†	0.41	0.00	0.44	0.17
SJC$\left(\tau^U,\tau^L\right)$	0.45	0.20	0.50†	0.45	0.20	0.46	0.27
Mixed Normal(ρ_1,ρ_2)	0.95	0.05	0.50	0.00	0.00	0.40	0.40

Figures marked with '†' are based on simulations or numerical quadrature.

2 Copula–Based Models for Time Series

The application of copulas to time series modelling currently has two distinct branches. The first is the application to multivariate time series, where the focus is in modelling the joint distribution of some random vector, $\mathbf{X}_t=[X_{1t},X_{2t},...,X_{nt}]'$, conditional on some information set $\mathcal{F}_{t-1}$. (The information set is usually $\mathcal{F}_{t-1}=\sigma(\mathbf{X}_{t-j};j\geq 1)$, though this need not necessarily be the case.) This is an extension of some of the early applications of copulas in statistical modelling where the random vector of interest could be assumed to be independent and identically distributed (iid), see Clayton (1978) and Cook and Johnson (1981) for example. This application leads directly to the consideration of time-varying copulas.

The second application in time series is to consider the copula of a sequence of observations of a univariate time series, for example, to consider the joint distribution of $[X_t, X_{t+1}, ..., X_{t+n}]'$. This application leads us to consider Markov processes and general nonlinear time series models. We discuss each of these branches of time series applications of copulas below.

2.1 Copula–based models for multivariate time series

In this sub-section we consider the extension required to consider the conditional distribution of $\mathbf{X}_t$ given some information set $\mathcal{F}_{t-1}$. Patton (2006a) defined a "conditional copula" as a multivariate distribution of (possibly correlated) variables that are each distributed as $Uniform(0,1)$ conditional on $\mathcal{F}_{t-1}$. With this definition, it is then possible to consider an extension of Sklar's theorem to the time series case:

$$\begin{aligned} &\mathbf{F}_t(\mathbf{x}|\mathcal{F}_{t-1}) \qquad (4)\\ &= \mathbf{C}_t(F_{1,t}(x_1|\mathcal{F}_{t-1}), F_{2,t}\left(x_2|\mathcal{F}_{t-1}\right), ..., F_{n,t}\left(x_n|\mathcal{F}_{t-1}\right)|\mathcal{F}_{t-1}), \quad \forall\ \mathbf{x} \in \mathbb{R}^n\ , \end{aligned}$$

where $X_i|\mathcal{F}_{t-1} \sim F_{i,t}$ and $\mathbf{C}_t$ is the conditional copula of $\mathbf{X}_t$ given $\mathcal{F}_{t-1}$.

The key complication introduced when applying Sklar's theorem to conditional distributions is that the conditioning set, $\mathcal{F}_{t-1}$, must be the same for all marginal distributions and the copula. Fermanian and Wegkamp (2004) and Fermanian and Scaillet (2005) consider the implications of a failure to use the same information set, and define a 'conditional *pseudo* copula' to help study this case[4]. Failure to use the same information set for all components on the right-hand side of equation (4) will generally imply that the function on the left-hand side of equation (4) is *not* a valid conditional joint distribution function. See Patton (2006a) for an example of this failure.

It is often the case in financial applications, however, that some of the information contained in $\mathcal{F}_{t-1}$ is not relevant for all variables. For example, it might be that each variable depends on its own first lag, but not on the lags of any other variable. Define $\mathcal{F}_{i,t-1}$ as the smallest subset of $\mathcal{F}_{t-1}$ such that $X_{it}|\mathcal{F}_{i,t-1} \overset{\mathcal{D}}{=} X_{it}|\mathcal{F}_{t-1}$. With this it is possible to construct each marginal distribution model using only $\mathcal{F}_{i,t-1}$, which will likely differ across margins, and then use $\mathcal{F}_{t-1}$ for the copula, to obtain a valid conditional joint distribution. However, it must be stressed that in general the same information set must be used across all marginal distribution models and the copula model,

[4] The 'pseudo-copula' of Fermanian and Wegkamp (2004) is not to be confused with the 'quasi-copula' of Alsina, *et al.* (1993) and Genest, *et al.* (1999), which is used to characterize operations on distribution functions that cannot correspond to an operation on random variables.

before possibly reducing each of these models by eliminating variables that are not significant/important[5].

The consideration of conditional copulas leads naturally to the question of whether these exhibit significant changes through time. Conditional correlations between financial asset returns are known to fluctuate through time, see Andersen, *et al.* (2006) and Bauwens *et al.* (2006) for example, and so it is important to also allow for time-varying conditional copulas. Patton (2002, 2006a) allows for time variation in the conditional copula by allowing the parameter(s) of a given copula to vary through time in a manner analogous to a GARCH model for conditional variance (Engle (1982) and Bollerslev (1986)). Jondeau and Rockinger (2006) employ a similar strategy. Rodriguez (2007), on the other hand, considers a regime switching model for conditional copulas, in the spirit of Hamilton (1989). Chollete (2005), Garcia and Tsafack (2007), and Okimoto (2006) employ a similar modelling approach, with the latter author finding that the copula of equity returns during the low mean-high variance state is significantly asymmetric (with non-zero lower tail dependence) while the high mean-low volatility state has a more symmetric copula. Panchenko (2005b) considers a semi-parametric copula-based model of up to five assets, building on Chen and Fan (2006b), discussed below, where the marginal distributions are estimated nonparametrically and the conditional copula is specified to be Normal, with a correlation matrix that evolves according to the DCC specification of Engle (2002). Lee and Long (2005) combine copulas with multivariate GARCH models in an innovative way: they use copulas to construct flexible distributions for the residuals from a multivariate GARCH model, employing the GARCH model to capture the time-varying correlation, and the copula to capture any dependence remaining between the conditionally uncorrelated standardised residuals.

It is worth noting that, for some of the more complicated models above, it can be difficult to establish sufficient conditions for stationarity, which is generally required for standard estimation methods to apply, as discussed in Section 2.3 below. Results for general classes of *univariate* nonlinear processes are presented in Carrasco and Chen (2002) and Meitz and Saikkonen (2004), however similar results for the multivariate case are not yet available.

2.2 Copula–based models for univariate time series

In addition to describing the cross-sectional dependence between two or more time series, copulas can also be used to describe the dependence between observations from a given univariate time series, for example, by captur-

[5] For example, in Patton (2006a) I study the conditional joint distribution of the returns on the Deutsche mark/U.S. dollar and Japanese Yen/U.S. dollar exchange rates. In that application Granger-causality tests indicated that the marginal distributions depended only on lags of the "own" variable; lags of other variables were not significant.

ing the dependence between $[X_t, X_{t+1}, ..., X_{t+n}]'$. If the copula is invariant through time and satisfies a constraint on its multivariate marginals[6], and the marginal distributions are identical and also invariant through time, then this describes a stationary Markov process. The main benefit of this approach to univariate time series modelling is that the researcher is able to specify the unconditional (marginal) distribution of X_t separately from the time series dependence of X_t. For example, the six joint distributions plotted in Figure 1 could be used to generate a stationary first-order Markov process, with the marginal distribution of X_t being $N(0,1)$, and with various copulas describing the dependence between X_t and X_{t+1}. In Figure 2 I plot the conditional mean of X_{t+1} given $X_t = x$, along with the conditional mean ± 1.65 times the conditional standard deviation of X_{t+1} given $X_t = x$, for each of the six distributions from Figure 1. In the upper left panel is the familiar case of joint normality: a linear conditional mean and constant conditional variance. The other five panels generally display non-linear conditional mean and variance functions. In Figure 3 I plot the density of X_{t+1} conditional on $X_t = -2,\ 0$, and 2. Now in the upper left panel we see the familiar figure of Normal conditional densities, while in the other panels the conditional densities are non-Normal. Amongst other things, the figures for the Student's t and mixed Normal copulas emphasise that radial symmetry of the joint distribution (i.e., symmetry around both the main diagonal and the off-diagonal) is not sufficient for symmetry of the conditional marginal densities.

Darsow, *et al.* (1992) study first-order Markov processes based on copulas. They provide a condition equivalent to the Chapman-Kolmogorov equations for a stochastic process that focusses solely on the copulas of the variables in the process. Furthermore, the authors are able to provide a necessary and sufficient condition for a stochastic process to be Markov by placing conditions on the multivariate copulas of variables in the process (in contrast with the Chapman-Kolmogorov equations which are necessary but not sufficient conditions). Ibragimov (2005, 2006) extends the work of Darsow, *et al.* (1992) to higher-order Markov chains and provides several useful results, and a new class of copulas. Beare (2007) studies the weak dependence properties of Markov chains through the properties of their copulas and, amongst other things, shows that tail dependence in the copula may result in the Markov chain not satisfying standard mixing conditions. Gagliardini and Gouriéroux (2007b) propose and study copula-based time series models for durations, generalising the autoregressive conditional duration model of Engle and Russell (1998).

[6] For example, if $n = 3$, then it is required that the marginal joint distribution of the first and second arguments is identical to that of the second and third arguments. Similar conditions are required for $n > 3$.

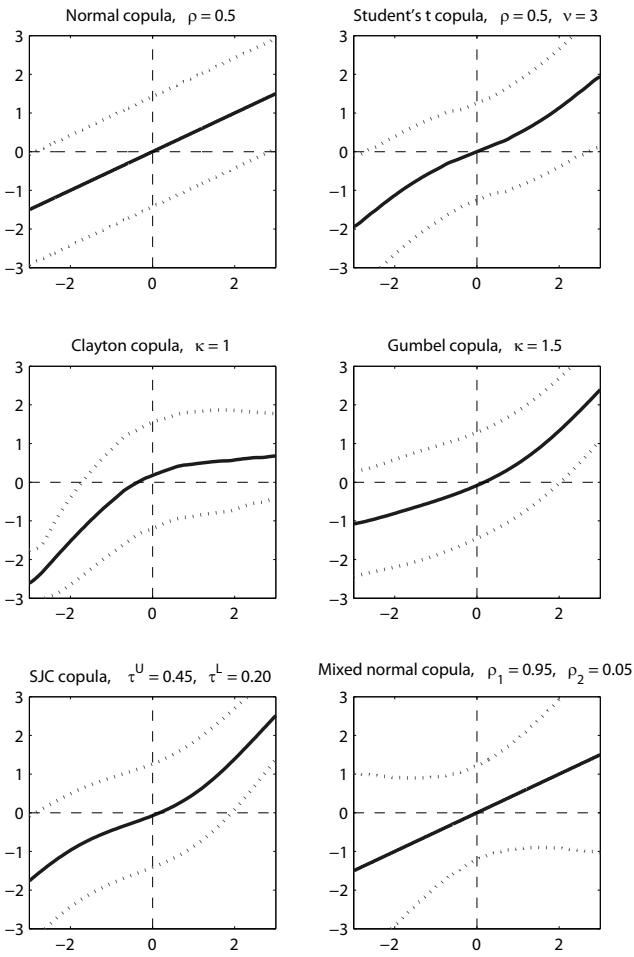

Fig. 2 Conditional mean ± 1.64 times the conditional standard deviation, for joint distributions with $N(0,1)$ marginal distributions and linear correlation coefficients of 0.5.

2.3 Estimation and evaluation of copula–based models for time series

The estimation of copula-based models for multivariate time series can be done in a variety of ways. For fully parametric models (the conditional marginal distributions and the conditional copula are all assumed known up to a finite-dimensional parameter) maximum likelihood (ML) is the ob-

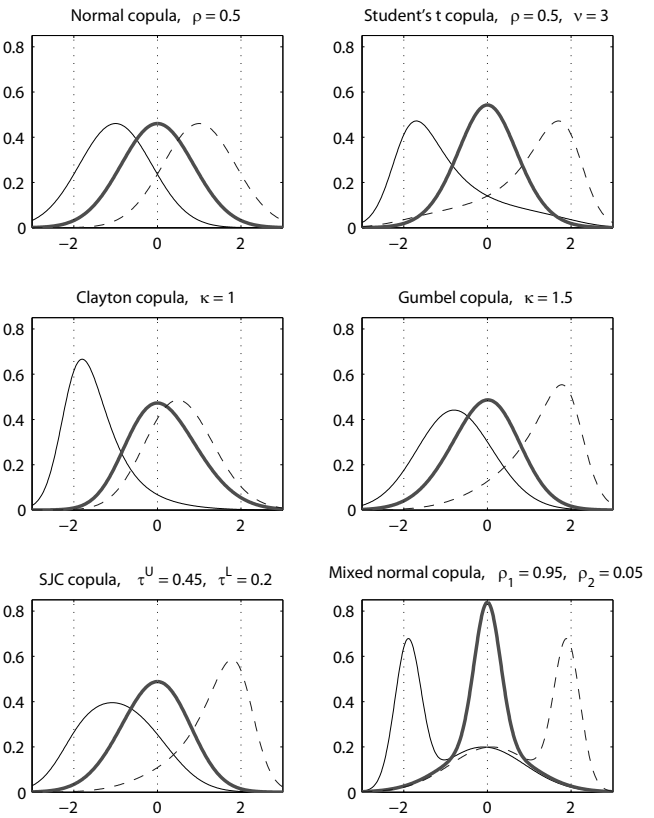

Fig. 3 Conditional densities of Y given $X = -2$ (thin line), $X = 0$ (thick line), $X = +2$ (dashed line), for joint distributions with $N(0,1)$ marginal distributions and linear correlation coefficients of 0.5.

vious approach. If the model is such that the parameters of the marginal distributions can be separated from each other and from those of the copula, then multi-stage ML estimation is an option. This method, sometimes known as the "inference functions for margins" (IFM) method, see Joe and Xu (1996) and Joe (1997, Chapter 10), involves estimating the parameters of the marginal distributions via univariate ML, and then estimating the parameter of the copula conditional on the estimated parameters for the marginal distributions. This estimation method has the benefit of being computationally tractable, at the cost of a loss of full efficiency. The theory for this estimation method for *iid* data is presented in Shih and Louis (1995) and Joe and Xu (1996). Patton (2006b) presents the theory for time series data, drawing

on the work on Newey and McFadden (1994) and White (1994), and some simulation results that motivate multi-stage estimation.

Fully nonparametric estimation methods for copula models in the *iid* case were studied by Genest and Rivest (1993) and Capéraà, *et al.,* (1997), amongst others. Fully nonparametric estimation of copulas for time series data was studied by Fermanian and Scaillet (2003) and Ibragimov (2005).

An important benefit of using copulas to construct multivariate models is that the models used in the marginal distributions need not be of the same type as the model used for the copula. One exciting possibility that this allows is non- or semi-parametric estimation of the marginal distributions, combined with parametric estimation of the copula. Such a model avoids the 'curse of dimensionality' by only estimating the one-dimensional marginal distributions nonparametrically, and then estimating the (multi-dimensional) copula parametrically. The theory for this estimator in the *iid* case is presented in Genest, *et al.* (1995) and Shih and Louis (1995). Theory for the time series case is presented in Chen and Fan (2006b) and Chen, *et al.* (2006). Chen and Fan (2006b) also consider the important case that the copula model may be mis-specified. Gagliardini and Gouriéroux (2007a) consider copula specifications that are semi-parametric, while Sancetta and Satchell (2004) consider semi-nonparametric copula models.

The estimation of fully parametric copula-based univariate time series models is discussed in Joe (1997, Chapter 8). Chen and Fan (2006a) consider the estimation of semi-parametric copula-based univariate time series models, where the unconditional distribution is estimated nonparametrically and the copula is estimated parametrically. The work of Ibragimov (2006) and Beare (2007) on conditions for some form of mixing to hold are also relevant here.

The evaluation of a given model is important in any econometric application, and copula-based modelling is of course no exception. The evaluation of copula-based models takes two broad approaches. The first approach evaluates the copula-based multivariate density model in its entirety, and thus requires methods for evaluating multivariate density models, see Diebold, *et al.* (1999) and Corradi and Swanson (2005). In the second approach one seeks to evaluate solely the copula model, treating the marginal distribution models as nuisance parameters. Fermanian (2005) and Scaillet (2007) consider such an approach for models based on *iid* data, while Malevergne and Sornette (2003) and Panchenko (2005a) consider tests for time series models. Genest *et al.* (2007) provide an extensive review of goodness-of-fit tests for copulas, focussing on the *iid* case, and present the results of a simulation study of the size and power of several tests.

Comparisons between a set of competing copula-based models can be done either via economic criteria, such in some of the papers reviewed in the next section, or statistical criteria. For the latter, likelihood ratio tests (either nested or, more commonly, non-nested, see Vuong (1989) and Rivers and Vuong (2002) for example) can often be used. Alternatively, information cri-

teria, such as the Akaike or Schwarz's Bayesian Information Criteria (AIC, BIC) can be used to penalise models with more parameters.

3 Applications of Copulas in Finance and Economics

The primary motivation for the use of copulas in finance comes from the growing body of empirical evidence that the dependence between many important asset returns is non-normal. One prominent example of non-normal dependence is where two asset returns exhibit greater correlation during market downturns than during market upturns. Evidence against the univariate normality of asset returns has a long history, starting with Mills (1927), but evidence against 'copula normality' has accumulated only more recently. Erb, *et al.* (1994), Longin and Solnik (2001) and Ang and Chen (2002), Ang and Bekaert (2002), Bae, *et al.* (2003) all document, without drawing on copula theory, evidence that asset returns exhibit non-normal dependence, that is, dependence that is not consistent with a Normal copula. This evidence has wide-ranging implications for financial decision-making, in risk management, multivariate option pricing, portfolio decisions, credit risk, and studies of 'contagion' between financial markets. In the remainder of this section I discuss some of the research done in these areas.

The first area of application of copulas in finance was risk management. Just as 'fat tails' or excess kurtosis in the distribution of a single random variable increases the likelihood of extreme events, the presence of non-zero tail dependence increases the likelihood of *joint* extreme events. As illustrated in Table 1, even copulas that are constrained to generate the same degree of linear correlation can exhibit very different dependence in or near the tails. The focus of risk managers on Value-at-Risk (VaR), and other measures designed to estimate the probability of 'large' losses, makes the presence of non-normal dependence of great potential concern. Cherubini and Luciano (2001), Embrechts, *et al.* (2003) and Embrechts and Höing (2006) study the VaR of portfolios using copula methods. Hull and White (1998) is an early paper on VaR for collections of non-normal variables. Rosenberg and Schuermann (2006) use copulas to consider 'integrated' risk management problems, where market, credit and operational risks must be considered jointly. McNeil, *et al.* (2005) and Alexander (2008) provide clear and detailed textbook treatments of copulas and risk management.

In derivatives markets non-normal dependence has key pricing, and therefore trading, implications. Any contract with two or more 'underlying' assets will generally have a price that is affected by both the strength and the shape of the dependence between the assets. A simple such contract is one that pays £1 if all underlying assets have prices above a certain threshold on the contract maturity date; another common contract is one that has a pay-off based on the minimum (or the maximum) of the prices of the un-

derlying assets on the contract maturity date. Even derivatives with just a single underlying asset may require copula methods if the risk of default by the counter-party to the contract is considered economically significant: these are so-called "vulnerable options". A recent book by Cherubini, *et al.* (2004) considers derivative pricing using copulas in great detail, and they provide an interesting introduction to copulas based on option pricing, as an alternative to the more standard statistical introductions in Joe (1997) and Nelsen (2006) for example. Other papers that consider option pricing with copulas include Rosenberg (2003), Bennett and Kennedy (2004), van den Goorbergh, *et al.* (2005) and Salmon and Schleicher (2006). Other authors, see Taylor and Wang (2004) and Hurd, *et al.* (2005), have instead used observed derivatives prices to find the implied copula of the underlying assets.

The booming market in credit derivatives (credit default swaps and collateralised debt obligations, for example) and the fact that these assets routinely involve multiple underlying sources of risks has lead to great interest in copulas for credit risk applications. An early contribution is from Li (2000), who was first to use copulas in a credit risk application, and was more generally one of the first to apply copulas in finance. See also Frey and McNeil (2001), Schönbucher and Schubert (2001) and Giesecke (2004) for applications to default risk. Duffie (2004) argues that copulas are too restrictive for certain credit risk applications.

One of the most obvious places where the dependence between risky assets impacts on financial decisions, and indeed was the example used at the start of this survey, is in portfolio decisions. Under quadratic utility and/or multivariate Normality (or more generally, multivariate ellipticality, see Chamberlain, 1983) the optimal portfolio weights depend only upon the first two moments of the assets under consideration, and so linear correlation adequately summarises the necessary dependence information required for an optimal portfolio decision. However when the joint distribution of asset returns is not elliptical, as the empirical literature cited above suggests, and when utility is not quadratic in wealth, the optimal portfolio weights will generally require a specification of the entire conditional distribution of returns. Patton (2004) considers a bivariate equity portfolio problem using copulas, and Garcia and Tsafack (2007) consider portfolio decisions involving four assets: stocks and bonds in two countries. The extension to consider portfolio decisions with larger numbers of assets remains an open problem.

The final broad topic that has received attention from finance researchers using copula methods is the study of financial 'contagion'. Financial contagion is a phenomenon whereby crises, somehow defined, that occur in one market lead to problems in other markets *beyond* what would be expected on the basis of fundamental linkages between the markets. The Asian crisis of 1997 is one widely-cited example of possible contagion. The difficulty in contagion research is that a baseline level of dependence between the markets must be established before it can be asserted that the dependence increased during a period of crisis. The heavy focus on levels and changes in depen-

dence has lead several researchers to apply copula methods in their study of contagion. Rodriguez (2007) was the first to apply copulas to contagion, which he studies with a Markov switching copula model. See Chollete, *et al.* (2005) and Arakelian and Dellaportas (2005) for alternative approaches.

Finally, there are a number of interesting papers using copulas in applications that do not fit into the broad categories discussed above. Bouyé and Salmon (2002) use copulas for quantile regressions, Breymann, *et al.* (2003) study the copulas of financial assets using intra-daily data, sampled at different frequencies, Daul, *et al.* (2003) and Demarta and McNeil (2005) study the Student's t copula and some useful extensions, Heinen and Rengifo (2003) use copulas to model multivariate time series of counts, Smith (2003) uses copulas to model sample selection, related to earlier work touching on copulas for this problem by Lee (1983), Bonhomme and Robin (2004) use copulas to model a large panel of earnings data, Bartram, *et al.* (2006) use a time-varying conditional copula model to study financial market integration between seventeen European stock market indices, Granger, *et al.* (2006) use copulas to provide a definition of a 'common factor in distribution', Hu (2006) uses mixtures of copulas to separate the degree of dependence from the 'shape' of dependence, and Brendstrup and Paarsch (2007) use copulas in a semiparametric study of auctions.

4 Conclusions and Areas for Future Research

In this survey I have briefly discussed some of the extensions of standard copula theory that are required for their application to time series modelling, and reviewed the existing literature on copula-based models of financial time series. This is a fast-growing field and the list of references will no doubt need updating in the near future.

In reviewing the extant literature on copulas for finance a number of topics stand out as possible avenues for future research. The most obvious, and perhaps difficult, is the extension of copula-based multivariate time series models to high dimensions. Existing models are not well-designed for higher-dimension applications; what is needed is a flexible yet parsimonious way of characterising high dimension copulas. A similar problem was faced in the multivariate ARCH literature in the mid-1990s, see Bauwens, *et al.* (2006). Two popular approaches to solve that problem are factor-based ARCH models and extensions, see Alexander and Chibumba (1998) and van der Weide (2002) for example, and the DCC model of Engle (2002) and its extensions, see Cappiello, *et al.* (2006) for example. Perhaps similar approaches will prove fruitful in high-dimensional copula modelling.

Acknowledgement I would particularly like to thank B. Beare, P. Embrechts, J.-D. Fermanian, T. Mikosch and J. Rosenberg for detailed comments and suggestions on this

chapter. I would also like to thank Y. Fan, J.-P. Kreiss, J. C. Rodriguez, C. Schleicher and T. Schuermann for helpful comments. Some Matlab code for copulas is available from http://www.economics.ox.ac.uk/members/andrew.patton/code.html.

References

Alexander, C. 2008: *Market Risk Analysis, Volume III*. Wiley & Sons, London, forthcoming.

Alexander, C. and Chibumba, A. (1997): Multivariate Orthogonal Factor GARCH. *Mimeo, University of Sussex*.

Alsina, C., Nelsen, R.B. and Schweizer, B. (1993): On the characterization of a class of binary operations on distribution functions. *Statistics and Probability Letters* **17**, 85–89.

Andersen, T.G., Bollerslev, T., Christoffersen, P.F. and Diebold, F.X. (2006): Volatility and Correlation Forecasting. *In: Elliott, G., Granger, C.W.J. and Timmermann, A. (Eds.): The Handbook of Economic Forecasting*. North Holland, Amsterdam.

Ang, A. and Bekaert, G. (2002): International Asset Allocation with Regime Shifts. *Review of Financial Studies* **15**, 1137–1187.

Ang, A. and Chen, J. (2002): Asymmetric Correlations of Equity Portfolios. *Journal of Financial Economics* **63**, 443–494.

Arakelian, V. and Dellaportas, P. (2005): Contagion tests via copula threshold models. *Mimeo, University of Crete*.

Bartram, S.M., Taylor, S.J. and Wang, Y.-H. (2006): The euro and European financial market dependence. *Journal of Banking and Finance* forthcoming.

Bae, K.-H., Karolyi, G.A. and Stulz, R.M. (2003): A New Approach to Measuring Financial Contagion. *Review of Financial Studies* **16**, 717–764.

Bauwens, L., Laurent, S. and Rombouts, J. (2006): Multivariate GARCH Models: A Survey. *Journal of Applied Econometrics* **21**, 79–109.

Beare, B. (2007): Copula-based mixing conditions for Markov chains. *Mimeo, University of Oxford*.

Bennett, M.N. and Kennedy, J.E. (2004): Quanto Pricing with Copulas. *Journal of Derivatives* **12**, 26–45.

Bollerslev, T. (1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bonhomme, S. and Robin, J.-M. (2004): Modeling Individual Earnings Trajectories using Copulas with an Application to the Study of Earnings Inequality: France, 1990-2002. *Mimeo, Université de Paris 1*.

Bouyé, E. and Salmon, M. (2002): Dynamic Copula Quantile Regressions and Tail Area Dynamic Dependence in Forex Markets. *Mimeo, University of Warwick*.

Brendstrup, B. and Paarsch, H.J. (2007): Semiparametric Identification and Estimation in Multi-Object English Auctions. *Journal of Econometrics* **141**, 84–108.

Breymann, W., Dias, A. and Embrechts, P. (2003): Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance* **3**, 1–16.

Capéraà, P., Fougères, A.-L. and Genest, C. (1997): A Nonparametric Estimation Procedure for Bivariate Extreme Value Copulas. *Biometrika* **84**, 567–577.

Cappiello, L., Engle, R.F. and Sheppard, K. (2003): Evidence of Asymmetric Effects in the Dynamics of International Equity and Bond Return Covariance. *Journal of Financial Econometrics* forthcoming.

Carrasco M. and Chen X. (2002): Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17–39.

Casella, G. and Berger, R.L. (1990): *Statistical Inference* Duxbury Press, U.S.A.

Chamberlain, G. (1983): A characterization of the distributions that imply mean-variance utility functions. *Journal of Economic Theory* **29**, 185–201.

Chen, X. and Fan, Y. (2006a): Estimation of copula-based semiparametric time series models. *Journal of Econometrics* **130**, 307–335.

Chen, X. and Fan, Y. (2006b): Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics* **135**, 125–154.

Chen, X., Fan, Y. and Tsyrennikov, V. (2006): Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* **101**, 1228–1240.

Cherubini, U. and Luciano, E. (2001): Value at Risk trade-off and capital allocation with copulas. *Economic Notes* **30**, 235–256.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004): *Copula Methods in Finance* John Wiley & Sons, England.

Chollete, L. (2005): Frequent extreme events? A dynamic copula approach. *Mimeo, Norwegian School of Economics and Business.*

Chollete, L., de la Peña, V. and Lu, C.-C. (2005): Comovement of international financial markets. *Mimeo, Norwegian School of Economics and Business.*

Christoffersen, P. (2008): Value–at–Risk models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 752–766. Springer Verlag, New York.

Clayton, D.G. (1978): A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.

Coles, S., Heffernan, J. and Tawn, J. (1999): Dependence measures for extreme value analyses. *Extremes* **2**, 339–365.

Cook, R.D. and Johnson, M.E. (1981): A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society* **43**, 210–218.

Corradi V. and Swanson, N.R. (2005): Predictive Density Evaluation. *In: Elliott, G., Granger, C.W.J. and Timmermann, A. (Eds.): Handbook of Economic Forecasting.* North Holland, Amsterdam.

Darsow, W.F., Nguyen, B. and Olsen, E.T. (1992): Copulas and Markov processes. *Illinois Journal of Mathematics* **36**, 600–642.

Daul, S., De Giorgi, E., Lindskog, F. and McNeil, A. (2003): The grouped t-copula with an application to credit risk. *RISK* **16**, 73–76.

Demarta, S. and McNeil, A.J. (2005): The t copula and related copulas. *International Statistical Review* **73**, 111–129.

Denuit, M. and Lambert, P. (2005): Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* **93**, 40–57.

Diebold, F.X., Gunther, T. and Tay, A.S. (1998): Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review* **39**, 863–883.

Diebold, F.X., Hahn, J. and Tay, A.S. (1999): Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High Frequency Returns on Foreign Exchange. *Review of Economics and Statistics* **81**, 661–673.

Duffie, D. (2004): *Clarendon Lecture in Finance, mimeo* Stanford University. http://www.finance.ox.ac.uk/NR/rdonlyres/9A26FC79-980F-4114-8033-B73899EADE88/0/slides_duffie_clarendon_3.pdf

Embrechts, P. and Höing, A. (2006): Extreme VaR scenarios in higher dimensions. *Mimeo ETH Zürich.*

Embrechts, P., McNeil, A. and Straumann, D. (2002): Correlation and Dependence Properties in Risk Management: Properties and Pitfalls. *In: Dempster, M. (Ed.): Risk Management: Value at Risk and Beyond.* Cambridge University Press.

Embrechts, P., Höing, A. and Juri, A. (2003): Using Copulae to bound the Value-at-Risk for functions of dependent risks. *Finance & Stochastics* **7**, 145–167.

Embrechts, P., Furrer, H. and Kaufmann, R. (2008): Different Kinds of Risk. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 729–751. Springer Verlag, New York.

Engle, R.F. (1982): Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation. *Econometrica* **50**, 987–1007.

Engle, R.F. (2002): Dynamic Conditional Correlation - A Simple Class of Multivariate GARCH Models. *Journal of Business and Economic Statistics* **20**, 339–350.

Engle, R.F. and Russell, J.R. (1998): Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica* **66**, 1127–1162.

Erb, C.B., Harvey, C.R. and Viskanta, T.E. (1994): Forecasting International Equity Correlations. *Financial Analysts Journal* **50**, 32–45.

Fermanian, J.-D. (2005): Goodness of fit tests for copulas. *Journal of Multivariate Analysis* **95**, 119–152.

Fermanian, J.-D. and Scaillet, O. (2003): Nonparametric estimation of copulas for time series. *Journal of Risk* **5**, 25–54.

Fermanian, J.-D. and Scaillet, O. (2005): Some statistical pitfalls in copula modeling for financial applications. *In: Klein, E. (Ed.): Capital Formation, Governance and Banking*. Nova Science Publishing.

Fermanian, J.-D. and Wegkamp, M. (2004): Time dependent copulas. *Mimeo, CREST*.

Fisher, R.A. (1932): *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Frey, R. and McNeil, A.J. (2001): *Modelling dependent defaults*. ETH, Zürich, E-Collection, http://e-collection.ethbib.ethz.ch/show?type=bericht&nr=273

Gagliardini, P. and Gouriéroux, C. (2007a): An Efficient Nonparametric Estimator for Models with Non-linear Dependence. *Journal of Econometrics* **137**, 187–229.

Gagliardini, P. and Gouriéroux, C. (2007b): Duration Time Series Models with Proportional Hazard. *Journal of Time Series Analysis* forthcoming.

Galambos, J. (1978): *The Asymptotic Theory of Extreme Order Statistics*. Wiley, New York.

Garcia, R. and Tsafack, G. (2007): Dependence Structure and Extreme Comovements in International Equity and Bond Markets. *Working paper, Université de Montreal*.

Genest, C. and Rivest, L.-P. (1993): Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association* **88**, 1034–1043.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995): A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* **82**, 543–552.

Genest, C., Quasada Molina, J.J., Rodríguez Lallena, J.A. and Sempi, C. (1999): A characterization of quasi-copulas. *Journal of Multivariate Analysis* **69**, 193–205.

Genest, C., Rémillard, B. and Beaudoin, D. (2007): Goodness-of-Fit Tests for Copulas: A Review and Power Study. *Insurance: Mathematics and Economics* forthcoming.

Giesecke, K. (2004): Correlated Default with Incomplete Information. *Journal of Banking and Finance* **28**, 1521–1545.

Grammig, J., Heinen, A. and Rengifo, E. (2004): An analysis of the submission of orders on Xetra, using multivariate count data. *CORE Discussion Paper 2004/58*.

Granger, C.W.J., Teräsvirta, T. and Patton, A.J. (2006): Common factors in conditional distributions for bivariate time series. *Journal of Econometrics* **132**, 43–57.

Hamilton, J.D. (1989): A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **57**, 357–384.

Heinen, A. and Rengifo, E. (2003): Modelling Multivariate Time Series of Count Data Using Copulas. *CORE Discussion Paper 2003/25*.

Hu, L. (2006): Dependence Patterns across Financial Markets: A Mixed Copula Approach. *Applied Financial Economics* **16**, 717–729.

Hull, J. and White, A. (1998): Value–at–Risk when daily changes in market variables are not normally distributed. *Journal of Derivatives* **5**, 9–19.

Hurd, M., Salmon, M. and Schleicher, C. (2005): Using copulas to construct bivariate foreign exchange distributions with an application to the Sterling exchange rate index. *Mimeo, Bank of England*.

Ibragimov, R. (2005): Copula-based dependence characterizations and modeling for time series. *Harvard Institute of Economic Research Discussion Paper* **2094**.

Ibragimov, R. (2006): Copula-based characterizations and higher-order Markov processes. *Mimeo, Department of Economics, Harvard University*.

Joe, H. (1997): *Multivariate Models and Dependence Concepts. Monographs in Statistics and Probability* **73**. Chapman and Hall, London.

Joe, H. and Xu, J.J. (1996): The Estimation Method of Inference Functions for Margins for Multivariate Models. *Working paper, Department of Statistics, University of British Columbia*.

Jondeau, E. and Rockinger, M. (2006): The copula-GARCH model of conditional dependencies: an international stock market application. *Journal of International Money and Finance* **25**, 827–853.

Lee, L.-F. (1983): Generalized econometric models with selectivity *Econometrica* **51**, 507–512.

Lee, T.-H. and Long, X. (2005): Copula-based multivariate GARCH model with uncorrelated dependent standardized returns. *Journal of Econometrics* forthcoming.

Li, D.X. (2000): On default correlation: a copula function approach. *Journal of Fixed Income* **9**, 43–54.

Longin, F. and Solnik, B. (2001): Extreme Correlation of International Equity Markets. *Journal of Finance* **56**, 649–676.

Malevergne, Y. and Sornette, D. (2003): Testing the Gaussian Copula Hypothesis for Financial Assets Dependences. *Quantitative Finance* **3**, 231–250.

McNeil, A.J., Frey, R. and Embrechts, P. (2005): *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, New Jersey.

Meitz M. and Saikkonen P. (2004): Ergodicity, mixing, and the existence of moments of a class of Markov models with applications to GARCH and ACD models. *Econometric Theory* forthcoming.

Mikosch, T. (2006): Copulas: Tales and Facts, with discussion and rejoinder. *Extremes* **9**, 3–62.

Miller, D.J. and Liu, W.-H. (2002): On the recovery of joint distributions from limited information. *Journal of Econometrics* **107**, 259–274.

Mills, F.C. (1927): *The Behavior of Prices*. National Bureau of Economic Research, New York.

Nelsen, R.B. (2006): *An Introduction to Copulas*, second Edition. Springer, New York.

Newey, W.K. and McFadden, D. (1994): Large Sample Estimation and Hypothesis Testing. *In: Engle, R.F. and McFadden, D. (Eds.): Handbook of Econometrics* **4**. North-Holland, Amsterdam.

Okimoto, T. (2006): New evidence of asymmetric dependence structure in international equity markets: further asymmetry in bear markets. *Journal of Financial and Quantitative Analysis* forthcoming.

Panchenko, V. (2005a): Goodness-of-fit Tests for Copulas. *Physica A* **355**, 176–182.

Panchenko, V. (2005b): Estimating and evaluating the predictive abilities of semiparametric multivariate models with application to risk management. *Mimeo, University of Amsterdam*.

Patton, A.J. (2002): *Applications of Copula Theory in Financial Econometrics*. Unpublished Ph.D. dissertation, University of California, San Diego.

Patton, A.J. (2004): On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation. *Journal of Financial Econometrics* **2**, 130–168.

Patton, A.J. (2006a): Modelling Asymmetric Exchange Rate Dependence. *International Economic Review* **47**, 527–556.

Patton, A.J. (2006b): Estimation of Multivariate Models for Time Series of Possibly Different Lengths. *Journal of Applied Econometrics* **21**, 147–173.

Rivers, D. and Vuong, Q. (2002): Model Selection Tests for Nonlinear Dynamic Models. *The Econometrics Journal* **5**, 1–39.

Rodriguez, J.C. (2007): Measuring financial contagion: a copula approach. *Journal of Empirical Finance* **14**, 401–423.

Rosenberg, J.V. (2003): Nonparametric pricing of multivariate contingent claims. *Journal of Derivatives* **10**, 9–26.

Rosenberg, J.V. and Schuermann, T. (2006): A general approach to integrated risk management with skewed, fat-tailed risks. *Journal of Financial Economics* **79**, 569–614.

Salmon, M. and Schleicher, C. (2006): Pricing Multivariate Currency Options with Copulas. *In: Rank, J. (Ed.): Copulas: From Theory to Application in Finance*. Risk Books, London.

Sancetta, A. and Satchell, S. (2004): The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory* **20**, 535–562.

Scaillet, O. (2007): Kernel based goodness-of-fit tests for copulas with fixed smoothing parameters. *Journal of Multivariate Analysis* **98**, 533–543.

Schönbucher, P. and Schubert, D. (2001): Copula Dependent Default Risk in Intensity Models. *Mimeo, Bonn University*.

Shih, J.H. and Louis, T.A. (1995): Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics* **51**, 1384–1399.

Silvennoinen, A. and Teräsvirta, T. (2008): Multivariate GARCH Models. *In: Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 201–229. Springer, New York.

Sklar, A. (1959): Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l' Institut Statistique de l'Universite de Paris* **8**, 229–231.

Smith, M.D. (2003): Modelling sample selection using Archimedean copulas. *Econometrics Journal* **6**, 99–123.

Taylor, S.J. and Wang, Y.-H. (2004): Option prices and risk-neutral densities for currency cross-rates. *Mimeo, Department of Accounting and Finance, Lancaster University*.

van den Goorbergh, R.W.J., C. Genest and Werker, B.J.M. (2005): Multivariate Option Pricing Using Dynamic Copula Models. *Insurance: Mathematics and Economics* **37**, 101–114.

van der Weide, R. (2002): GO-GARCH: A Multivariate Generalized Orthogonal GARCH Model. *Journal of Applied Econometrics* **17**, 549–564.

Vuong, Q. (1989): Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57**, 307–333.

White, H. (1994): *Estimation, Inference and Specification Analysis. Econometric Society Monographs* **22**, Cambridge University Press, Cambridge, U.K.

Credit Risk Modeling

David Lando

Abstract The chapter gives a broad outline of the central themes of credit risk modeling starting with the modeling of default probabilities, ratings and recovery. We present the two main frameworks for pricing credit risky instruments and credit derivatives. The key credit derivative - the Credit Default Swap - is introduced. The premium on this contract provides a meausure of the credit spread of the reference issuer. We then provide some key empirical works looking at credit spreads thorugh CDS contracts and bonds and finish with a description of the role of correlation in credit risk modeling.

1 Introduction

Credit risk modeling is a rapidly growing area of financial economics and financial engineering. Banks and other financial institutions are applying increasingly sophisticated methods for assessing the risk of their loan portfolios and their counterparty exposure on derivatives contracts. The market for credit derivatives is growing at an extreme pace with the credit default swap and the CDO markets as the primary sources of growth. These new markets and better data availability on the traditional corporate bond market have provided new laboratories for financial economists to test asset pricing theories, to look at capital structure decisions, and to understand financial innovation.

Classical credit risk analysis is concerned with deciding whether a loan should be granted and, after a loan has been granted, trying to assess the risk of default. Modern credit risk analysis still addresses these issues but there is more focus on pricing of loans and corporate bonds in secondary markets

David Lando
Copenhagen Business School, Department of Finance, Solbjerg Plads 3, DK-2000 Frederiksberg, Denmark, e-mail: dl.fi@cbs.dk

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_35,

and on the pricing and hedging of derivative contracts whose pay-offs depend on the performance of a single bond or loan or on a portfolio of bonds or loans. This survey will focus on the modeling of corporate bond prices and credit spreads and on some implications for modeling credit derivatives. These are areas of credit risk modeling where rich data sets and well-developed pricing models allow for extensive econometric analysis. There are interesting econometric challenges in more traditional credit scoring but we only have time to touch briefly upon this topic. Other topics will not be covered at all, including the role of credit risk in the macro-economy (credit crunches, the role of credit in the propagation of business cycles), implications for optimal capital structure decisions of firms and capital adequacy requirements for banks.

2 Modeling the Probability of Default and Recovery

The decision theoretic problem of whether to grant a loan or not has been attacked for a long time using financial ratios but the systematic attempts of looking at which predictors perform well was started in Altman (1968) and Beaver (1966) who mainly rely on discriminant analysis, an attempt of classifying firms into defaulting or non-defaulting groups based on company characteristics. The outcome of the analysis is a discriminant function which maps the relevant predictors into a single score and classifies the company as a defaulter or a non-defaulter based on whether the score is above or below a certain threshold. The derivation of the discriminant function and the threshold can be based on a likelihood approach or on a decision theoretic approach, in which a cost is assigned to misclassification. For the precise arguments, see for example Anderson (1984). This method of analysis was in part chosen for computational convenience and it is not easily adapted to a dynamic framework with time-varying covariates. The method is also not well-suited for handling the effects of covariates that are common to all firms such as business cycle indicators. The use of computers have greatly facilitated the use of logistic regression methods, see for example Shumway (2001) and survival analytic methods based on hazard regressions, see for example Duffie et al. (2004) which both deal easily with dynamic features of the model, common covariates and which offer default probability predictions useful for risk management.

As an alternative to developing full models for default probabilities of individual firms, investors in corporate bonds often use the ratings which are assigned to bond issuers by the major rating agencies. The ratings are based on analysis of the company's key ratios and on meetings with the issuers. Ratings play important roles in a regulatory context when assigning risk-weights to assets, in defining loan covenants and investment guidelines and as a signalling device to less informed investors. Given these important roles, there

is a substantial literature on the statistical behavior of ratings focusing on various forms of non-Markov behavior, business cycle dependence, the ability of ratings to forecast default and their information content measured through price reactions in markets around rating changes. A number of references to this literature can be found in Lando (2005)

The recovery rates on defaulted bonds play an important role in pricing both bonds and derivatives but systematic research into what determines recovery rates is still relatively new. For a comprehensive recent study, see Acharya et al. (2007) which also contains a number of references.

3 Two Modeling Frameworks

Models for default probabilities and recovery are important inputs to pricing models for loans, bonds and derivatives but we need more structure in the models than what we get from a statistical model to actually perform the pricing. A statistical model typically gives us estimates of default probabilities as a function of certain explanatory variables, but the dynamics of the explanatory variables are not modeled. A pricing model needs to include the full dynamics of the explanatory variables.

There are broadly speaking two approaches to pricing corporate bonds and these approaches are outlined in this section. One approach - sometimes referred to as the structural approach - views equity and bonds (and other claims to the firm's assets) as derivative contracts on the market value of a firm's assets. It then uses option pricing theory to price these claims. This approach is systematically carried out in Merton (1974) in a Black-Scholes setting. To briefly summarize the approach, assume that the market value of a firm's assets follows a geometric Brownian motion

$$dV_t = \mu V_t dt + \sigma V_t dW_t$$

where W is a standard Browninan motion and that the firm has issued a zero coupon bond promising to pay the principal D at the maturity date T. Assume that the actual pay-off of the bond at maturity is $\min(V_T, D)$. This means that debt holders receive their payment in full if the firm has sufficient assets to pay, but if assets are insufficient the bondholders take over the remaining assets. In the setting of a Black-Scholes market with a constant interest rate r the price of this bond is give at time zero as

$$B_0 = D\exp(-rT) - P^{BS}(V_0, D, \sigma, r, T)$$

where P^{BS} denotes the value of a European put option in the Black-Scholes setting. We can translate this bond price into a (promised) yield

$$y(T) = -\frac{1}{T} \log \frac{B_0}{D}$$

and by varying T in this expression we obtain what is known as the risk structure of interest rates. We obtain the credit spread by subtracting the riskless rate r from the promised yield. The most important observation from this relationship is that yield spreads increase when leverage, i.e. the ratio of D to V, increases and when asset volatility increases. Note that credit spreads are increasing in volatility and do not distinguish between whether the volatility is systematic or non-systematic. The expected return of the bond is sensitive to whether volatility risk is systematic or not. So one should carefully distinguish between credit spreads and expected returns on corporate bonds.

There are many modifications and extensions of the fundamental setup outlined above. Bonds may pay continuous or lumpy coupons, there may be a lower boundary which when reached by the asset value of the firm causes the firm to default. This lower boundary could represent bond safety covenants or liquidity problems of the firm and it could represent future capital structure decisions of the firm trying to maintain a stationary leverage ratio. Dynamics of key inputs may be changed as well, allowing for example for stochastic interest rates or jumps in firm asset value. For a survey, see Lando (2005). The structural models differ in how literally the option approach is taken. Several of the models impose default boundaries exogenously and obtain prices on coupon bonds by summing the prices of zero-coupon bonds obtained from the risk structure of interest rates - an approach which does not make sense in the Merton model, see Lando (2005) for a simple explanation.

A special branch of the literature endogenizes the default boundary, see Leland (1994) and Leland and Toft (1996), or takes into account the ability of owners to act strategically by exploiting that the costs of bankruptcy are borne by the debt holders, see for example Anderson and Sundaresan (1996).

Applying contingent claims analysis to corporate bonds differs from contingent claims pricing in many important respects. Most notably, the market value of the underlying assets is not observable. In applications the asset value must therefore be estimated and for companies with liquid stocks trading a popular approach has been to estimate the underlying asset value and the drift and volatility of the asset value by looking at equity prices as transformations (by the Black-Scholes call option formula) of the asset value. It is therefore possible using the transformation theorem from statistics to write down the likelihood function for the observed equity prices and to estimate the volatility and the drift along with the value of the underlying assets. For more on this, see Duan (1994).

The option-based approach is essential for looking at relative prices of different liabilities in a firm's capital structure, for discussing optimal capital structure in a dynamic setting and for defining market-based predictors of default.

The second approach to spread modeling - sometimes referred to as the reduced-form approach or the intensity-based approach - takes as given a

stochastic intensity of default which plays the same role for default risk as the short rate plays in government term structure modeling. The advantage of this approach is precisely that it integrates the modeling of corporate bonds with modeling of the default-free term structure of interest rates, and this makes it particularly suitable for econometric specification of the evolution of credit spreads and for pricing credit derivatives.

Early intensity models are in Jarrow and Turnbull (1995) and Madan and Unal (1998) and the full integration of stochastic intensities in term structure models is found in Lando (1994, 1998) and Duffie and Singleton (1999). The integration using a Cox process setup proceeds as follows. A non-negative stochastic process λ describes the instantaneous probability of default under a risk-neutral measure Q, i.e.

$$Q(\tau \in (t, t + \Delta t] | \mathcal{F}_t) = 1_{\{\tau > t\}} \lambda_t \Delta t$$

where $\mathcal{F}_t$ contains information at time t including whether the default time τ is smaller than t. Assume that the intensity process λ and a process for the riskless short rate r are adapted to a filtration $(\mathcal{G}_t)$ where $\mathcal{G}_t \subset \mathcal{F}_t$, and assume that the default time τ of a firm is modeled as

$$\tau = \inf \left\{ t : \int_0^t \lambda(s) ds \geq E_1 \right\}.$$

where E_1 is an exponentially distributed random variable with mean 1 which is independent of the filtration $(\mathcal{G}_t)$. In accordance with the formal definition of an intensity, it can be shown that $1_{\{\tau \leq t\}} - \int_0^t \lambda(s) 1_{\{\tau \geq s\}} ds$ is an $\mathcal{F}_t$-martingale. The key link to term structure modeling can be seen most easily from the price $v(0, T)$ of a zero coupon bond maturing at T with zero recovery in default (in contrast with B_0 defined earlier) and issued by a company with default intensity λ under the risk-neutral measure Q :

$$\begin{aligned} v(0, T) &= E^Q \left[\exp \left(- \int_0^T r_s \, ds \right) 1_{\{\tau > T\}} \right] \\ &= E^Q \left[\exp \left(- \int_0^T (r_s + \lambda_s) ds \right) \right]. \end{aligned}$$

A key advantage of the intensity models is that the functional form for the price of the defaultable bond is the same as that of a default-free zero-coupon bond in term structure modeling, and therefore the machinery of affine processes can be applied. This is true also when the formula is extended to allow for recovery by bond holders in the event of default, see for example Lando (2005).

In the reduced-form setting prices of coupon bonds are typically computed by summing the prices of zero-coupon bonds - an approach which clearly should be applied with caution. It is suitable for valuing credit default swaps

(see below) in which the contract has no influence on the capital structure of the firm or for assessing the default risk of small additional exposures of a firm for example in the context of counterparty risk in derivatives contracts. Pricing large new corporate bond issues will require an intensity which takes into account the effects on leverage of the new issue.

The division between the two approaches should not be taken too literally. A structural model with incomplete information under certain assumptions can be recast as an intensity model as shown in Duffie and Lando (2001), and nothing precludes an intensity model from letting the default intensity depend on the asset value of the firm and other firm specific variables.

Estimation of intensity models is performed in Duffee (1999) and Driessen (2005) using a Kalman filter approach. Duffie and Singleton (1997) focus on swap rates in an intensity-based setting. The default intensity process is treated as a latent process similar to the short-rate process in classical term structure modeling. This means that estimation proceeds in close analogue with estimation of term structure models for default-free bonds. There is one important difference, however, in how risk premia are specified for default intensities and for the riskless rate. To understand the difference in a single factor setting, think of a diffusion-driven short rate depending on a single Brownian motion, as for example in the Cox-Ingersoll-Ross setting. The change of measure between the physical measure and the risk-neutral measure corresponds to a change of drift of the Brownian motion. A similar change of measure can be made for the intensity controlling the default time of a defaultable bond issuer. However, it is also possible to have an intensity process which is changed by a strictly positive, multiplicative factor when going from the physical to the risk-neutral measure. This gives rise to an important distinction between compensation for variation in default risk and compensation for jump-event risk. For more on this distinction, see Jarrow et al. (2005) and for a paper estimating the two risk premia components separately, see Driessen (2005).

4 Credit Default Swap Spreads

The first empirical work on estimating intensity models for credit spreads employed corporate bond data, but this is rapidly changing with the explosive growth of the credit default swap (CDS) market. The CDS contracts have become the benchmark for measuring credit spreads, at least for the largest corporate issuers. A CDS is a contract between two parties: one who buys and one who sells protection against default of a particular bond issuer which we call the *reference* issuer. The protection buyer pays a periodic premium, the CDS premium, until whichever comes first, the default event of the reference issuer or the maturity date of the contract. In the event of default of the reference issuer before the maturity of the CDS contract, the protection seller

compensates the protection buyer for the loss on a corporate bond issued by the reference firm. The compensation is made either by paying a cash amount equivalent to the difference between the face value and the post default market value of the defaulted bond or by paying the face value while taking physical delivery of the defaulted bond. In practice, there is a delivery option allowing the protection buyer to deliver one of several defaulted bonds with pre-specified characteristics in terms of seniority and maturity. The CDS premium is set such that the initial value of the contract is zero.

Forming a portfolio of a credit risky bond and a CDS contract with the same maturity protecting against default on that bond, one has a position close to a riskless bond and this gives an intuitive argument for why the credit default swap premium ought to be close to a par bond spread on a corporate bond. For a more rigorous argument, see Duffie (1999).

It is illustrative to use the intensity setting to compute the fair premium on a CDS, i.e. the premium which gives the contract value 0 at initiation. In practice, this relationship is primarily used to infer the default intensity from observed CDS prices and then price other derivatives from that intensity, or to analyze risk premia of default by comparing market implied default intensities with actual default intensities.

To clearly illustrate the principle, consider a stylized CDS with maturity T years, where premium payments are made annually. As above, let the default intensity of the reference issuer be denoted λ under a risk-neutral measure Q. Then the present value of the CDS premium payments made by the protection buyer before the issuer defaults (or maturity) is simply

$$\widetilde{\pi}^{pb} = c \sum_{t=1}^{T} E^Q \exp(-\int_0^t r_s \, ds) 1_{\{\tau > t\}} = c \sum_{t=1}^{T} v(0,t)$$

where $v(0,t)$ is the value of a zero recovery bond issued by the reference firm which we used as the basic example above. If we write the probability of surviving past t under the risk neutral measure as

$$S(0,t) = E^Q \big[\exp(-\int_0^t \lambda_s \, ds) \big]$$

and we assume (as is commonly done when pricing CDS contracts) that the riskless rate and the default intensity are independent, then we can express the value of the protection payment made before default occurs as

$$\widetilde{\pi}^{pb} = c \sum_{t=1}^{T} p(0,t) S(0,t)$$

where $p(0,t)$ is the value of a zero-coupon riskless bond maturing at date t. This formula ignores the fact that if a default occurs between two payment dates, the protection buyer pays a premium determined by the fraction of a

period that elapsed from the last premium payment to the default time. If we assume that default happens in the middle between two coupon dates, then the value of this extra payment should be added to get the value of the protection payment:

$$\pi^{pb} = \widetilde{\pi}^{pb} + \frac{c}{2}\sum_{t=1}^{T} p(0, t-\frac{1}{2})\Big(S(0,t) - S(0,t-1)\Big).$$

The value of the protection seller's obligation is more complicated (but certainly manageable) if we insist on taking into account the exact timing of default. To simplify, however, we again assume that if a default occurs between two default dates, it occurs in the middle, and the settlement payment is made at that date. Furthermore, we assume a recovery per unit of principal equal to δ, so that the protection seller has to pay $1-\delta$ to the protection buyer per unit of notional. Still assuming independence of the riskless rate and the default intensity, we obtain

$$\pi^{ps} = (1-\delta)\sum_{t=1}^{T} p(0, t-\frac{1}{2})\Big(S(0,t) - S(0,t-1)\Big).$$

The fair CDS premium c can now be found by equating π^{pb} and π^{ps}. In practice payments are often made quarterly in rates equal to one fourth of the quoted annual CDS premium.

There are several advantages of using CDS contracts for default studies: First of all, they trade in a variety of maturities thus automatically providing a term structure for each underlying name. They are becoming very liquid for large corporate bond issuers, their documentation is becoming standardized and unlike corporate bonds they do not require a benchmark bond for extracting credit spreads.

Currently, there is an explosion of papers using CDS data. One early contribution is Blanco et al. (2005) who among other things study lead/lag relationships between CDS premia and corporate bond spreads on a sample of European investment grade names. They find evidence that CDS contracts lead corporate bonds. Longstaff et al. (2005) study the size of the CDS spread compared to corporate bond spreads measured with respect to different benchmark riskless rates. Assuming that the CDS premium represents pure credit risk they are then able to take out the credit risk component of corporate bonds and ask to what extent liquidity-related measures influence the residual spread on corporate bonds.

5 Corporate Bond Spreads and Bond Returns

The current price of a corporate bond in the Merton model - even under risk neutrality - is smaller than the price of a default-free bond. Therefore, the yield spread is positive even when there is risk neutrality. This part of the corporate bond spread is the expected loss component. Empirically, corporate spreads are found to be larger than the expected loss component, and in fact they are apparently so much larger that it is questioned whether reasonable assumptions on risk premia for default can explain the remainder. This is the essence of the credit risk puzzle. How can actual spreads be so much larger than the expected loss component? Roughly, potential contributions to the corporate spread can be divided into the expected loss component, components due to priced market risk factors in corporate bonds and other factors, such as taxes (in the US coupons on corporate bonds are taxed at the state level whereas Treasury bond coupons are not), a liquidity premium on corporate bonds, and the choice of riskless benchmark (is the Treasury rate the appropriate benchmark to use?). Early papers pointing to the difficulty of structural models in explaining spreads are Jones et al. (1984) and Sarig and Warga (1989). Newer papers attempting to decompose spreads in structural model settings along the dimensions mentioned above include Huang and Huang (2003) and Eom et al. (2003). Many papers study spreads in time series and/or regression frameworks where inspiration on relevant covariates to include come from form structural models, including Elton et al. (2001) and Collin-Dufresne et al. (2001). A consensus seems to be building that the Treasury rate is not an appropriate benchmark in US markets for defining a riskless rate and using a rate closer to the swap rate at least removes part of the credit risk puzzle. Using the swap rate as riskless benchmark also brings credit spreads measured from CDS contracts closer to spreads measured from corporate bonds as shown in Longstaff et al. (2005). Still, there is no definitive answer yet to what fraction of credit spreads are in fact explained by default-related factors.

6 Credit Risk Correlation

Dependence of default events is a key concern of regulators who are looking into issues of financial stability. Performing empirical studies on dependence based solely on events, similar to the basic default studies, is difficult since the number of defaults is fairly limited. Fortunately, a very large market of Collateralized Debt Obligations has emerged and the pricing of securities in this market is intimately linked to correlation. This therefore allows us to analyze market implied correlation. A CDO is an example of an asset backed security in which the collateral consists of loans or bonds, and where securities issued against the collateral are prioritized claims to the cash flow of

the collateral. The lowest priority claim, the so-called equity tranche, is most sensitive to default in the underlying pool and the highest priority claim, the senior tranche, is strongly protected against default. This closely resembles how equity, junior debt and senior debt are prioritized claims to a firm's assets, but there are often 5-7 different 'layers' in the CDO tranches.

There are roughly speaking three approaches to pricing CDOs.

1. In the copula based approach one takes as given the marginal default intensities of the individual loans in the collateral pool and defines a joint distribution of default by applying a copula function, i.e. a device for collecting a family of marginal distributions into a joint distribution preserving the marginal distributions. Often the loans are actually given the same default intensity and the copula function is then chosen from a parametric class of copulas attempting to fit the model to observed tranche prices. A first paper in this area is Li (2000) but see also Schönbucher (2003).
2. A full modeling approach is taken in Duffie and Gârleanu (2001) where the default intensities of individual issuers are given a factor structure, such that the individual intensities are a sum of a common 'factor' intensity and idiosyncratic intensities. While one can facilitate computations by working with affine intensity processes, the model still requires many parameters as inputs, and this is one reason why it is not adapted as quickly by market participants despite its more appealing structure. One may also question the implicit assumption of conditional independence whereby the only dependence that the defaults have is through their common dependence on the aggregate factor. Conditionally on this factor the defaults are independent. This rules out direct contagion among names in the portfolio.
3. In the third approach, the process of cumulative losses is modeled directly without focus on the individual issues. One can impose simple jumps to accommodate one default at a time or one can choose to work with possibilities of multiple defaults.

As noted, the choice of modeling here is linked to the issue of the extent to which defaults are correlated only through their common dependence on the economic environment or whether there is true contagion. This contagion could take several forms: The most direct is that defaults cause other defaults but it might also just be the case that defaults cause an increasing likelihood of default for other issuers. Das et al. (2004) test whether the correlation found in US corporate defaults can be attributed to common variation on the estimated intensity of default for all issuers. By transforming the timescale, they test the joint hypothesis that the intensities of default of individual firms are estimated correctly and that the defaults are conditionally independent given the realized intensities, and they reject the joint hypothesis. This is an indication that contagion events could be at play but the reason for the rejection could also lie in a mis-specification of the intensities, neglecting for

example an unobserved common default risk factor, sometimes referred to as a frailty factor in line with the survival analysis literature. An inaccurately specified underlying intensity could also be the problem. A proper understanding and parsimonious modeling of dependence effects in CDO pricing remains an important challenge.

References

Acharya, V., Bharath, S. and Srinivasan, A. (2007): Does industry-wide distress affect defaulted firms? evidence from creditor recoveries. *Journal of Financial Economics* to appear.

Altman, E. (1968): Financial ratios: Discriminant analysis, and the prediction of corporate bankruptcy. *Journal of Finance* **23**, 589–609.

Anderson, R. and Sundaresan, S. (1996): Design and valuation of debt contracts. *Review of Financial Studies* **9**, 37–68.

Anderson, T. W. (1984): *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York.

Beaver, W. (1966): Financial ratios and the prediction of failure. *Journal of Accounting Research. Supplement: Empirical research in accounting: Selected studies 1966* **4**, 77–111.

Blanco, R., Brennan, S. and Marsh, I. W. (2005): An empirical analysis of the dynamic relationship between investment grade bonds and credit default swaps. *Journal of Finance* **60**, 2255–2281.

Collin-Dufresne, P., Goldstein, R. S. and Martin, J. (2001): The determinants of credit spread changes. *Journal of Finance* **56**, 2177–2207.

Das, S., Duffie, D., Kapadia, N. and Saita, L. (2004): Common failings: How defaults are correlated. *Journal of Finance* to appear.

Driessen, J. (2005): Is Default Event Risk Priced in Corporate Bonds? *Review of Financial Studies* **18**, 165–195.

Duan, J. (1994): Maximum likelihood estimation using price data of the derivatives contract. *Mathematical Finance* **4**, 155–167.

Duffee, G. (1999): Estimating the price of default risk. *Review of Financial Studies* **12**, 197–226.

Duffie, D. (1999): Credit swap valuation. *Financial Analysts Journal* **55**, 73–87.

Duffie, D. and Gârleanu, N. (2001): Risk and valuation of collateralized debt obligations. *Financial Analysts Journal* **57**, 41–59.

Duffie, D. and Lando, D. (2001): Term structures of credit spreads with incomplete accounting information. *Econometrica* **69**, 633–664.

Duffie, D., Saita, L. and Wang, K. (2004): Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics* to appear.

Duffie, D. and Singleton, K. (1997): An econometric model of the term structure of interest-rate swap yields. *The Journal of Finance* **52**, 1287–1321.

Duffie, D. and Singleton, K. (1999): Modeling term structures of defaultable bonds. *Review of Financial Studies* **12**, 687–720.

Elton, E. J., Gruber, M. J., Agrawal, D. and Mann, C. (2001): Explaining the rate spread on corporate bonds. *Journal of Finance* **56**, 247–277.

Eom, Y., Helwege, J. and Huang, J.-Z. (2003): Structural Models of Corporate Bond Pricing. *Review of Financial Studies* **17**, 499–544.

Huang, J. and Huang, M. (2003): How Much of the Corporate-Treasury Yield Spread is due to Credit Risk? *Working Paper, Stanford University.*

Jarrow, R., Lando, D. and Yu, F. (2005): Default risk and diversification: Theory and applications. *Mathematical Finance* **15**, 1–26.

Jarrow, R. and Turnbull, S. (1995): Pricing options on financial securities subject to credit risk. *Journal of Finance* **50**, 53–85.

Jones, E., Mason, S. and Rosenfeld, E. (1984): Contingent claims analysis of corporate capital structures: An empirical investigation. *Journal of Finance* **39**, 611–625.

Lando, D. (1994): *Three essays on contingent claims pricing*. PhD Dissertation, Cornell University.

Lando, D. (1998): On Cox processes and credit risky securities. *Review of Derivatives Research* **2**, 99–120.

Lando, D. (2005): *Credit Risk Modeling - Theory and Applications*. Princeton University Press.

Leland, H. E. (1994): Corporate debt value, bond covenants, and optimal capital structure. *The Journal of Finance* **49**, 157–196.

Leland, H. E. and Toft, K. (1996): Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *Journal of Finance* **51**, 987–1019.

Li, D. (2000): On default correlation: A copula function approach. *The Journal of Fixed Income* **9**.

Longstaff, F. A., Mithal, S. and Neis, E. (2005): Corporate yield spreads: Default risk or liquidity? new evidence from the credit-default swap market. *Journal of Finance* **60**, 2213–2253.

Madan, D. and Unal, H. (1998): Pricing the risks of default. *Review of Derivatives Research* **2**, 121–160.

Merton, R. C. (1974): On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* **29**, 449–470.

Sarig, O. and Warga, A. (1989): Some empirical estimates of the risk structure of interest rates. *Journal of Finance* **46**, 1351–1360.

Schönbucher, P. (2003): *Credit Derivatives Pricing Models: Models, Pricing and Implementation*. Wiley, New York.

Shumway, T. (2001): Forecasting bankruptcy more efficiently: A simple hazard model. *Journal of Business* **74**, 101–124.

Evaluating Volatility and Correlation Forecasts

Andrew J. Patton and Kevin Sheppard *

Abstract This chapter considers the problems of evaluation and comparison of volatility forecasts, both univariate (variance) and multivariate (covariance matrix and/or correlation). We pay explicit attention to the fact that the object of interest in these applications is unobservable, even ex post, and so the evaluation and comparison of volatility forecasts often rely on the use of a "volatility proxy", i.e. an observable variable that is related to the latent variable of interest. We focus on methods that are robust to the presence of measurement error in the volatility proxy, and to the conditional distribution of returns.

1 Introduction

This chapter considers the problems of evaluation and comparison of univariate and multivariate volatility forecasts, with explicit attention paid to the fact that in such applications the object of interest is unobservable, even *ex post*. Thus the evaluation and comparison of volatility forecasts must rely on direct or indirect methods of overcoming this difficulty. Direct methods use a "volatility proxy", i.e. some observable variable that is related to the latent

Andrew J. Patton
Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, United Kingdom, e-mail: andrew.patton@economics.ox.ac.uk

Kevin Sheppard
Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, United Kingdom, e-mail: kevin.sheppard@economics.ox.ac.uk

* We thank Torben Andersen, Rob Engle, Neil Shephard and especially Peter Christoffersen for helpful comments, and Runquan Chen for excellent research assistance. The first author gratefully acknowledges financial support from the Leverhulme Trust under grant F/0004/AF.

DOI: 10.1007/978-3-540-71297-8_36,

variable of interest. We will assume the existence of an unbiased volatility proxy, such as daily squared returns for the daily conditional variance of returns. Indirect methods of overcoming the latent nature of the variable of interest include comparing forecasts via mean-variance portfolio decisions or comparisons based on portfolio "tracking error".

A guiding theme of our analysis is the focus on tests that are "robust" in two ways: Firstly, we want tests that are robust to the presence of noise in the volatility proxy, if a proxy is used. The presence of noise in the proxy may affect – indeed, is likely to affect – the power of the tests, but it should not distort the asymptotic size of the test. Secondly, we desire tests that are robust to the conditional distribution of returns. Some existing volatility forecast evaluation and comparison techniques rely, in an important way, on assumptions about the conditional distribution of returns beyond the conditional second moment. While these techniques may have strong economic motivations, and thus be of interest in their own right, we argue that they are not pure tests of *volatility* forecasts and that they should not be used without a convincing economic argument. Such arguments will generally be specific to each application.

The main results and recommendations of this chapter, based on theoretical considerations and on simulation studies, can be summarised as follows. Firstly, we suggest a minor modification of the widely-used Mincer-Zarnowitz regression for testing volatility forecast optimality which exploits the additional structure that holds under the null hypothesis. This "MZ-GLS" test has good size and much better finite sample power than other MZ tests. Secondly, we find that the use of loss functions that are "non-robust", in the sense of Patton (2006), can yield perverse rankings of forecasts, even when accurate volatility proxies are employed. This emphasises the need to pay careful attention to the selection of the loss function in Diebold and Mariano (1995) and West (1996) tests when evaluating volatility forecasts. Amongst the class of robust loss functions for volatility forecast evaluation, and the multivariate generalisation of these loss functions provided in this chapter, our simulations suggest that the "QLIKE" loss function yields the greatest power. Finally, consistent with the large and growing literature on realised volatility, our simulations clearly demonstrate the value of higher-precision volatility proxies, such as realised variance or daily high-low range, see Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2004). Even simple estimators based on 30-minute returns provide large gains in power and improvements in finite-sample size.

The problems which arise in evaluating a single volatility forecast, a set of volatility forecasts, or a complete covariance matrix are so similar that dividing the chapter into univariate and multivariate sections would be misleading. This chapter avoids the false dichotomy between univariate and multivariate volatility forecasting and evaluation as much as possible. As a results, this chapter is primarily organised along the method of evaluation, not the dimension of the problem studied. Sections 2 and 3 of this chapter focus on direct

methods for forecast evaluation and comparison using a volatility proxy, while Sect. 4 discusses indirect, or economic, methods.

1.1 Notation

We will present the notation for the general multivariate case first and specialise the notation for the univariate case where needed. Let $\mathbf{r}_t \equiv [r_{1t},\ r_{2t}, \ldots, r_{Kt}]'$ be the $K \times 1$ vector valued variable whose conditional covariance is of interest. The information set used in defining the conditional variance is denoted $\mathcal{F}_{t-1}$, and is assumed to contain the history of past returns, but may also include other variables and/or variables measured at a higher frequency than $\mathbf{r}_t$, such as intra-daily returns. Denote $V[\mathbf{r}_t|\mathcal{F}_{t-1}] \equiv V_{t-1}[\mathbf{r}_t] \equiv \mathbf{\Sigma}_t$, a $K \times K$ symmetric, positive definite matrix composed of elements $\sigma_{ij,t}$ where $\sigma_{ii,t} \equiv \sigma^2_{i,t}$ denotes the conditional variance of the i^{th} return and $\sigma_{ij,t}$ denotes the conditional covariance between the i^{th} and j^{th} series. We will assume throughout that $E[\mathbf{r}_t|\mathcal{F}_{t-1}] \equiv E_{t-1}[\mathbf{r}_t] = \mathbf{0}$, and thus $\mathbf{\Sigma}_t = E_{t-1}[\mathbf{r}_t\mathbf{r}_t']$. Let $\boldsymbol{\varepsilon}_t \equiv \mathbf{\Sigma}_t^{-1/2}\mathbf{r}_t$ denote the "standardised vector of returns", where $\mathbf{\Sigma}_t^{1/2}$ is a matrix that satisfies $\mathbf{\Sigma}_t^{1/2\prime}\mathbf{\Sigma}_t^{1/2} = \mathbf{\Sigma}_t$.[2] We assume that

$$\mathbf{r}_t|\mathcal{F}_{t-1} \sim \mathbf{F}_t(\mathbf{0}, \mathbf{\Sigma}_t) \tag{1}$$

where $\mathbf{F}_t$ is some distribution with zero mean and finite covariance $\mathbf{\Sigma}_t$. In some applications we will use a stronger condition that $\mathbf{r}_t|\mathcal{F}_{t-1}$ has a constant conditional distribution and hence constants higher order moments, i.e.

$$\mathbf{r}_t|\mathcal{F}_{t-1} \sim \mathbf{F}(\mathbf{0}, \mathbf{\Sigma}_t)\ . \tag{2}$$

Let a forecast of the conditional covariance of $\mathbf{r}_t$ be denoted $\mathbf{H}_t$, or $\mathbf{H}_t^A$, $\mathbf{H}_t^B$, $\mathbf{H}_t^C$, ... if there multiple forecasts are under analysis. The loss function in the multivariate case is $L : \mathbb{M}_+^K \times \mathcal{H}^K \to \mathbb{R}_+$, where the first argument of L is $\mathbf{\Sigma}_t$ or some proxy for $\mathbf{\Sigma}_t$, denoted $\widehat{\mathbf{\Sigma}}_t$, and the second is $\mathbf{H}_t$. $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the non-negative and positive real line, $\mathcal{H}^K$ is a compact subset of $\mathbb{M}_{++}^K$, and $\mathbb{M}_+^K$ and $\mathbb{M}_{++}^K$ denote the positive semi-definite and positive definite subspaces of the set of all real symmetric $K \times K$ matrices. Note that $\mathbb{M}_+^1 = \mathbb{R}_+$ and $\mathbb{M}_{++}^1 = \mathbb{R}_{++}$.

Commonly used univariate volatility proxies are the squared return, r_t^2, realised volatility based on m intra-daily observations ("m-sample RV"), $RV_t^{(m)}$, and the range, RG_t, while commonly used covariance proxies are the outer

[2] For example, the "square root" matrix, $\mathbf{\Sigma}_t^{1/2}$, can be based on the Cholesky or the spectral decomposition of $\mathbf{\Sigma}_t$. The Cholesky square-root is not invariant to the order of the variables in $\mathbf{r}_t$, and so any subsequent computations may change following a simple re-ordering of these variables. For this reason we recommend the use of the square-root based on the spectral decomposition.

product of returns, $\mathbf{r}_t\mathbf{r}_t'$ and realised covariance, $RC_t^{(m)}$. In this chapter we will treat the forecasts as "primitive", and make no attempt to study them via the models, if any, that are used to generate the forecasts. Specification, estimation and selection of univariate and multivariate volatility models are considered in further chapters of this volume, cf. Chib (2008), Koopman (2008), Silvennoinen and Teräsvirta (2008), Teräsvirta (2008) and Zivot (2008).

We define a conditionally unbiased volatility proxy, denoted $\widehat{\sigma}_t^2$, and a conditionally unbiased covariance proxy, denoted $\widehat{\mathbf{\Sigma}}_t = [\widehat{\sigma}_{ij}]_t$ as variables that satisfy:

$$E\left[\widehat{\sigma}_t^2|\mathcal{F}_{t-1}\right] = \sigma_t^2 \quad a.s., \; t = 1, 2, .. \tag{3}$$

$$E\left[\widehat{\mathbf{\Sigma}}_t|\mathcal{F}_{t-1}\right] = \mathbf{\Sigma}_t \quad a.s., \; t = 1, 2, .. \tag{4}$$

We will assume that at least one such proxy is available in all cases, though we make no further assumptions about the accuracy or consistency of the proxy.

2 Direct Evaluation of Volatility Forecasts

In this section we review tests for the evaluation of volatility forecasts using a volatility proxy. Recalling that $\sigma_t^2 \equiv V\left[r_t|\mathcal{F}_{t-1}\right]$ and drawing on Definition 5.2 White (1996), we define an optimal univariate volatility forecast as one that satisfies the following null hypothesis:

$$\begin{aligned} H_0^* &: h_t = \sigma_t^2 \; a.s., \; t = 1, 2, \ldots \\ \text{vs.} \quad H_1^* &: h_t \neq \sigma_t^2 \text{ for some } t \end{aligned} \tag{5}$$

The corresponding null for multivariate volatility forecasts is:

$$\begin{aligned} H_0^* &: \mathbf{H}_t = \mathbf{\Sigma}_t \; a.s., \; t = 1, 2, \ldots \\ \text{vs.} \quad H_1^* &: \mathbf{H}_t \neq \mathbf{\Sigma}_t \text{ for some } t \end{aligned} \tag{6}$$

The above null hypotheses are the ones that would ideally be tested in forecast evaluation tests. Instead, simple implications of these hypotheses are usually tested; we review the most common tests below.

2.1 Forecast optimality tests for univariate volatility forecasts

One common method of evaluating forecasts is the Mincer-Zarnowitz, or MZ, regression (cf. Mincer and Zarnowitz (1969)) , which involves regressing the realisation of a variable on its forecast (see also Theil (1958)). However, unlike standard forecast evaluation problems, the conditional variance is never observed, and the usual MZ regression is infeasible for volatility forecast evaluation. Using a conditionally unbiased estimator of the conditional variance, the feasible MZ regression:

$$\widehat{\sigma}_t^2 = \alpha + \beta h_t + e_t \tag{7}$$

yields unbiased estimates of α and β. The usual MZ test can then be conducted:

$$\begin{aligned} H_0 &: \alpha = 0 \cap \beta = 1 \\ \text{vs.} \quad H_1 &: \alpha \neq 0 \cup \beta \neq 1 \end{aligned} \tag{8}$$

The OLS parameter estimates will be less accurately estimated the larger the variance of $\left(\sigma_t^2 - \widehat{\sigma}_t^2\right)$, which suggests the use of high frequency data to construct more accurate volatility proxies, cf. Andersen and Bollerslev (1998). While using less accurate estimates of $\widehat{\sigma}_t^2$ affects the precision of α and β, and thus the power of the test to detect deviations from forecast optimality, it does not affect the validity of the test.[3]

The standard MZ regression can detect certain deviations from H_0^* but is not consistent against all possible deviations. While it is possible to construct consistent MZ test (see Bierens (1990), de Jong (1996) Bierens and Ploberger (1997)), so-called "augmented MZ regressions", constructed using additional $\mathcal{F}_{t-1}$-measurable instruments to increase the power of the test in certain directions, are more common. Standard instruments include the lagged volatility proxy, $\widehat{\sigma}_{t-1}^2$, the lagged standardised volatility proxy, $\widehat{\sigma}_{t-1}^2/h_{t-1}$, sign based indicators, $I_{[r_{t-1}<0]}$, or combinations of these. These instruments are motivated by a desire to detect neglected nonlinearities or persistence in the volatility forecast. Grouping the set of $\mathcal{F}_{t-1}$-measurable instruments into a vector $\mathbf{z}_{t-1}$, the augmented MZ regression, null and alternative hypotheses are

$$\begin{aligned} \widehat{\sigma}_t^2 &= \alpha + \beta h_t + \boldsymbol{\gamma}' \mathbf{z}_{t-1} + e_t \\ H_0 \ &: \ \alpha = 0 \cap \beta = 1 \cap \boldsymbol{\gamma} = \mathbf{0} \\ \text{vs.} \quad H_1 \ &: \ \alpha \neq 0 \cup \beta \neq 1 \cup \boldsymbol{\gamma} \neq \mathbf{0} \ . \end{aligned} \tag{9}$$

[3] Chernov (2007) provides a detailed analysis of the implications of measurement error in the *regressors*, such as the $\mathbf{z}_{t-1}$ variables in (9).

Note that the residual of the above regression will generally be heteroskedastic, even under H_0^*, and so robust standard errors White (1980) are required. This prompted some authors to consider a different approach: testing for serial correlation in the standardised volatility proxy:

$$\begin{aligned} \frac{\widehat{\sigma}_t^2}{h_t} &= \delta + \vartheta \frac{\widehat{\sigma}_{t-1}^2}{h_{t-1}} + u_t \\ H_0 &: \delta = 1 \cap \vartheta = 0 \\ \text{vs. } H_1 &: \delta \neq 1 \cup \vartheta \neq 0 \end{aligned} \tag{10}$$

This approach generates residuals that are homoskedastic under H_0^* if the noise in the proxy $\left(\eta_t \equiv \widehat{\sigma}_t^2/\sigma_t^2\right)$ has constant conditional variance[4], and so if this assumption holds then robust standard errors are not required.

2.2 MZ regressions on transformations of $\widehat{\sigma}_t^2$

The use of squared returns in MZ regressions has caused some researchers concern, as statistical inference relies on fourth powers of the returns, and thus returns that are large in magnitude have a large impact on test results. One frequently proposed alternative is to use transformations of the volatility proxy and forecast to reduce the impact of large returns (see Jorion (1995), Bollerslev and Wright (2001)). Two such examples are:

$$|r_t| = \alpha + \beta\sqrt{h_t} + e_t, \text{ and} \tag{11}$$

$$\log\left(r_t^2\right) = \alpha + \beta \log\left(h_t\right) + e_t \tag{12}$$

Using these regressions can result in size distortions, even asymptotically, due to noise in the volatility proxy, $\widehat{\sigma}_t^2$. Take the regression in (11) as an example: under H_0^* the population values of the OLS parameter estimates are easily shown to be:

$$\begin{aligned} \alpha &= 0 \\ \beta &= \begin{cases} E\left[|\varepsilon_t|\right], & \text{if } E_{t-1}\left[|\varepsilon_t|\right] \text{ is constant} \\ \sqrt{\frac{\nu-2}{\pi}}\Gamma\left(\frac{\nu-1}{2}\right)/\Gamma\left(\frac{\nu}{2}\right), & \text{if } r_t|\mathcal{F}_{t-1} \sim \textit{Student's } t\left(0,\sigma_t^2,\nu\right),\ \nu > 2 \\ \sqrt{2/\pi} \approx 0.80, & \text{if } r_t|\mathcal{F}_{t-1} \sim N\left(0,\sigma_t^2\right) \end{cases} \end{aligned}$$

where *Student's* $t\left(0,\sigma_t^2,\nu\right)$ is a Student's t distribution with mean zero, variance σ_t^2 and ν degrees of freedom. When returns have the Student's t distribution, the population value for β decreases towards zero as $\nu \downarrow 2$, indicating that excess kurtosis in returns increases the distortion in this parameter. In

[4] This corresponds to conditional homokurticity, discussed below, of the returns if the volatility proxy used is a squared return.

the regression in (12) the population OLS parameters under H_0^* are:

$$\alpha = \begin{cases} E\left[\log \varepsilon_t^2\right], \\ \log(\nu-2) + \Psi\left(\frac{1}{2}\right) - \Psi\left(\frac{\nu}{2}\right), \text{ if } r_t|\mathcal{F}_{t-1} \sim \textit{Student's } t\left(0,\sigma_t^2,\nu\right), \nu > 2 \\ -\log(2) - \gamma_E \approx -1.27, \qquad \text{if } r_t|\mathcal{F}_{t-1} \sim N\left(0,\sigma_t^2\right) \end{cases}$$

$$\beta = 1$$

where Ψ is the digamma function and $\gamma_E = -\Psi(1) \approx 0.58$ is Euler's constant, cf. Harvey et al. (1994). Under the Student's t distribution, the above expression shows $\alpha \to -\infty$ as $\nu \downarrow 2$. Thus while both of these alternative MZ regressions may initially appear reasonable, without some modification they lead to the undesirable outcome that the perfect volatility forecast, $h_t = \sigma_t^2$ *a.s.*, will be rejected with probability approaching one as the sample size increases. In both cases the perverse outcomes are the result of the imperfect nature of any volatility proxy; if volatility was observable, regressions on the transformation would lead to the correct result.

A second alternative is to adjust the volatility proxy, either exactly or approximately, so as to make it unbiased for the quantity of interest, σ_t or $\log \sigma_t$, and thus avoid any asymptotic size distortions, see Bollerslev and Wright (2001), Christodoulakis and Satchell (2004), Andersen et al. (2005). However, these adjustments require further assumptions about the distribution of the noise in the volatility proxy, and test statistics may be misleading if the assumptions are violated.

2.3 Forecast optimality tests for multivariate volatility forecasts

The results for testing optimality of conditional volatility can be directly extended to conditional covariance forecasts. The simplest optimality test examines the unique elements of the forecast covariance $\mathbf{H}_t$ separately using feasible MZ regressions

$$\widehat{\sigma}_{ij,t} = \alpha_{ij} + \beta_{ij} h_{ij,t} + e_{ij,t} \tag{13}$$

or augmented MZ regressions

$$\widehat{\sigma}_{ij,t} = \alpha_{ij} + \beta_{ij} h_{ij,t} + \boldsymbol{\gamma}_{ij}' \mathbf{z}_{ij,t} + e_{ij,t} \,. \tag{14}$$

resulting in $K(K+1)/2$ regressions and test statistics. This may be problematic, particularly when K is large as some rejections are expected even if the conditional covariance is correct. Alternatively, a joint test, that all of the coefficients are simultaneously zero can be tested the by forming a vector

process using the half-vector operator[5] (vech),

$$\text{vech}(\widehat{\boldsymbol{\Sigma}}_t) = \boldsymbol{\alpha} + \text{diag}(\boldsymbol{\beta})\mathbf{H}_t + \boldsymbol{\varepsilon}_t \tag{15}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $K(K+1)/2$ parameter vectors and diag is the diagonal operator[6]. Explicit expressions for a heteroskedasticity consistent covariance estimator can be found in many panel data texts (see, e.g. Arellano (2003)). Despite availability of a joint test, the finite sample properties may be adversely affected by the large dimension of the parameter covariance matrix. A simple specification could be constructed using only a common parameter across all pairs,

$$\text{vech}(\widehat{\boldsymbol{\Sigma}}_t) = \alpha + \beta\text{vech}\left(\mathbf{H}_t\right) + \boldsymbol{\varepsilon}_t \tag{16}$$

and testing whether α and β are 0 and 1, respectively[7].

2.4 Improved MZ regressions using generalised least squares

The residuals from the feasible MZ and augmented MZ regressions above will generally be heteroskedastic, and the size and power properties of these tests can be improved using generalised least squares. Consider a decomposition of the volatility proxy into the true variance and a multiplicative error term: $\widehat{\sigma}_t^2 \equiv \sigma_t^2\eta_t$, where $E_{t-1}\left[\eta_t\right] = 1$, and the feasible univariate MZ regression from above:

$$\widehat{\sigma}_t^2 = \alpha + \beta h_t + e_t$$

Under H_0^*, residuals from this regression will be

$$\begin{aligned} e_t = \widehat{\sigma}_t^2 - h_t &= \sigma_t^2\left(\eta_t - 1\right) \\ \text{so } E_{t-1}\left[e_t\right] &= 0 \\ \text{but } \ V_{t-1}\left[e_t\right] &= \sigma_t^4 V_{t-1}\left[\eta_t\right] \equiv \varsigma_t^2 \end{aligned}$$

If ς_t^2 was known for all t, then GLS estimation of the feasible MZ regression would simply be

---

[5] This operator stacks the columns of the lower triangle of a square matrix (Magnus and Neudecker (2002)).

[6] The diagonal operator transforms a $K \times 1$ vector into a $K \times K$ matrix with the vector along the diagonal. That is, $\text{diag}(\boldsymbol{\beta}) = (\boldsymbol{\beta}\boldsymbol{\iota}') \odot \mathbf{I}_K$, where $\boldsymbol{\iota}$ is a $K \times 1$ vector of ones, $\mathbf{I}_K$ is the K-dimensional identity matrix, and $\odot$ represents the Hadamard product (for element-by-element multiplication).

[7] While the parameters of this regression can be estimated using OLS by stacking the elements of $\text{vech}(\widehat{\boldsymbol{\Sigma}}_t)$, the errors will generally be cross-sectionally correlated and White standard errors will not be consistent. Instead, a pooled-panel covariance estimator appropriate for cross-correlated heteroskedastic data should be used.

$$\frac{\widehat{\sigma}_t^2}{\varsigma_t} = \alpha \frac{1}{\varsigma_t} + \beta \frac{h_t}{\varsigma_t} + \widetilde{e}_t \ . \tag{17}$$

In the special case where the volatility proxy is a squared return, $\widehat{\sigma}_t^2 = r_t^2$ and where the standardised returns are *conditionally homokurtic*[8], i.e. $E_{t-1}\left[r_t^4\right]/\sigma_t^4 = \kappa \ \forall \ t$, the MZ-GLS takes a very simple form:

$$\frac{\widehat{\sigma}_t^2}{h_t} = \alpha \frac{1}{h_t} + \beta + \widetilde{e}_t \ . \tag{18}$$

since $V_{t-1}[\widehat{\sigma}_t^2] = \sigma_t^4 \kappa = h_t^2 \kappa \propto h_t^2$ under H_0^*.

This simple standardisation is not guaranteed to produce efficient estimates for arbitrary volatility proxies, for example realised variance or range. However, it is generally the case that the volatility of the proxy is increasing in the level of the proxy. Thus using the specification in (18) may result in improved finite sample performance, even when not fully efficient. A formal proof of this conjecture is left to future research[9].

The application of MZ-GLS to covariance forecast evaluation is similar although there are choices which may affect the finite sample properties of the parameter estimators. The direct extension of the univariate framework specifies the MZ-GLS as

$$\frac{\widehat{\sigma}_{ij,t}}{\varsigma_{ij,t}} = \alpha_{ij} \frac{1}{\varsigma_{ij,t}} + \beta_{ij} \frac{h_{ij,t}}{\varsigma_{ij,t}} + \widetilde{e}_{ij,t} \ . \tag{19}$$

where $\varsigma_{ij,t} \equiv V_{t-1}[\widehat{\sigma}_{ij,t}]$, and again $\alpha_{ij} = 0 \cap \beta_{ij} = 1$ under H_0^*. If the volatility proxies are squares or cross-products of conditionally homokurtic returns, the MZ-GLS can again be specified using the forecasts as weights

$$\frac{\widehat{\sigma}_{ij,t}}{\sqrt{h_{ii,t}h_{jj,t} + h_{ij,t}^2}} = \alpha_{ij} \frac{1}{\sqrt{h_{ii,t}h_{jj,t} + h_{ij,t}^2}} + \beta_{ij} \frac{h_{ij,t}}{\sqrt{h_{ii,t}h_{jj,t} + h_{ij,t}^2}} + \widetilde{e}_{ij,t} \tag{20}$$

The denominator of the left-hand-side of the above equation can be written as $\sqrt{h_{ii,t}h_{jj,t}}\sqrt{1 + \varrho_{ij,t}^2}$. It may be noted that the contribution to the het-

[8] This holds, for example, if returns are conditionally Normally distributed, or conditionally Student's t distributed with constant degrees of freedom greater than 4.

[9] In the presence of intra-daily heteroskedasticity, the variance of the proxy will not generally be proportional to the conditional variance and so a direct estimate will be required. If the proxy is realised variance, such an estimator will require an estimate of the "integrated quarticity", see Barndorff-Nielsen and Shephard (2004). For application of GLS, the estimator needs to be consistent in T, thus requiring that $m = O\left(T^{\delta}\right)$ for $\delta \geq 1$, with the specific rate δ depending on assumptions about the underlying diffusion. In finite samples, integrated quarticity is often estimated with substantial error, and so scaling by a consistent estimator of the proxy variance, while accurate asymptotically, may perform worse than simply scaling by the forecast.

eroskedasticity of $\widehat{\sigma}_{ij,t}$ from the $\sqrt{1+\varrho_{ij,t}^2}$ term is generally small since this is bounded between 1 and $\sqrt{2}$. A slightly simpler specification for evaluating conditional covariance models that accounts for the largest contributors to the heteroskedasticity of $\widehat{\sigma}_{ij,t}$ can be specified as

$$\frac{\widehat{\sigma}_{ij,t}}{\sqrt{h_{ii,t}h_{jj,t}}} = \alpha_{ij}\frac{1}{\sqrt{h_{ii,t}h_{jj,t}}} + \beta_{ij}\varrho_{ij,t} + \widetilde{e}_{ij,t} \; . \tag{21}$$

and can be interpreted as a regression of the correlation proxy on the forecast correlation[10], cf. Tse (2000). When $i = j$ all of these specifications reduce to the volatility MZ-GLS in (18).

2.5 Simulation study

To assess the size and power properties of the MZ and MZ-GLS tests, we compare variations on two methods for evaluating the performance of volatility and covariance forecasts. The first method (MZ1) involves regressing a volatility proxy (either squared daily returns, or a realised volatility proxy constructed to resemble a realised variance based on 30-minute or 5-minute returns) on a constant and the forecast:

$$\widehat{\sigma}_t^2 = \alpha + \beta h_t + e_t$$

and then testing

$$\begin{aligned} H_0 &: \alpha = 0 \cap \beta = 1 \\ \text{vs } H_1 &: \alpha \neq 0 \cup \beta \neq 1 \end{aligned}$$

This regression will have heteroskedastic errors under H_0^* and the covariance of the OLS parameters must be estimated using a heteroskedasticity consistent estimator, such as White's robust covariance estimator, cf. White (1980). Alternatively, as discussed above, the volatility forecasts themselves can be used to obtain homoskedastic regression residuals via GLS. In this case we use OLS to estimate

$$\frac{\widehat{\sigma}_t^2}{h_t} = \alpha\frac{1}{h_t} + \beta + \widetilde{e}_t$$

We denote these two approaches as MZ1 (White) and MZ1-GLS. The proxy used, either daily returns squared or realised variance is denoted $RV^{(m)}$, where m is the number of intra-daily returns used to construct the realised

[10] Strictly speaking, $\frac{\widehat{\sigma}_{ij,t}}{\sqrt{\widehat{\sigma}_{i,t}^2\widehat{\sigma}_{j,t}^2}}$ would be the correlation proxy, however this proxy is not generally unbiased since it is a non-linear function of unbiased estimators. The modification in (21) is a compromise that makes use of the volatility forecasts, $h_{ii,t}$ and $h_{jj,t}$, which are unbiased and error free under H_0^*.

variance. The second approach (MZ2) uses the standardised proxy in the regression, and has residuals that are homoskedastic under H_0^*, as long as the noise in the proxy $\left(\eta_t \equiv \widehat{\sigma}_t^2/\sigma_t^2\right)$ has constant conditional variance:

$$\frac{\widehat{\sigma}_t^2}{h_t} = \delta + \vartheta \frac{\widehat{\sigma}_{t-1}^2}{h_{t-1}} + u_t \tag{22}$$

and then tests

$$\begin{aligned} H_0 &: \delta = 1 \cap \vartheta = 0 \\ \text{vs } H_1 &: \delta \neq 1 \cup \vartheta \neq 0 \end{aligned}$$

While the MZ2 regression has an intuitive appeal, we are not aware of any studies of its performance relative to the MZ1 regressions in a realistic scenario.

The data generating process used in the simulation was specified as a standard GARCH(1,1)

$$DGP \ : \ r_t = \sigma_t \varepsilon_t, \ t = 1, 2, \ldots, T \tag{23}$$

$$\text{where } \ E_{t-1}\left[\varepsilon_t\right] = 0 \text{ and } \ E_{t-1}\left[\varepsilon_t^2\right] = 1$$

$$\sigma_t^2 = 0.05 + 0.85\sigma_{t-1}^2 + 0.10 r_{t-1}^2 \tag{24}$$

$$T = \{100, 250, 500, 1000\}$$

We specify the distribution of the standardised innovations as:

$$\begin{aligned} \varepsilon_t \ &= \ \sum_{m=1}^{78} \xi_{mt} \\ \xi_{mt} &\overset{\text{IID}}{\sim} \ N\left(0, 1/78\right) \end{aligned} \tag{25}$$

The distribution of the standardised innovations is designed to allow for DGP-consistent high-frequency estimators to be computed while preserving the GARCH(1,1) specification for the variance. The volatility proxies are computed from the ξ_{mt} as

$$RV_t^{(m)} = \sigma_t^2 \sum_{i=1}^{m} \left(\sum_{j=\lambda(i-1)+1}^{\lambda i} \xi_{it} \right)^2 \tag{26}$$

where $\lambda = 78/m$. We considered three values for m, the number of intra-daily returns. $m = 1$ corresponds to the use of daily squared returns as a proxy; $m = 13$ corresponds to using half-hourly returns for a stock traded on the NYSE; and $m = 78$ corresponds to using five-minute returns for a stock traded on the NYSE.

Size Comparison of MZ-type Tests for Volatility Forecasts

	T = 100			T = 250		
	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
MZ1 (White)	0.23	0.09	0.07	0.16	0.07	0.07
MZ1-GLS	0.11	0.07	0.06	0.08	0.06	0.05
MZ1-GLS (White)	0.15	0.07	0.06	0.10	0.06	0.06
MZ2	0.07	0.06	0.05	0.05	0.05	0.05
MZ2 (White)	0.15	0.07	0.07	0.10	0.06	0.06
	T = 500			T = 1000		
	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
MZ1 (White)	0.12	0.07	0.06	0.10	0.06	0.06
MZ1-GLS	0.06	0.05	0.05	0.06	0.05	0.05
MZ1-GLS (White)	0.07	0.05	0.05	0.06	0.05	0.05
MZ2	0.05	0.05	0.05	0.05	0.05	0.05
MZ2 (White)	0.08	0.05	0.06	0.07	0.05	0.05

Table 1 Rejection frequency of test statistics using a nominal size of 5% based on asymptotic critical values. All test statistics were computed using Wald tests using either White's heteroskedasticity-consistent covariance estimator or the standard OLS variance covariance formula as indicated. Three proxies were used for the latent volatility: the daily squared returns ($RV^{(1)}$), a 13-sample realised variance estimator ($RV^{(13)}$) and a 78-sample realised variance ($RV^{(78)}$).

The forecast model was designed to reflect the fact that most volatility forecasting models are able to closely match the unconditional variance. Forecasts predominantly differ in their specification for the dynamics of conditional variance. The model used is also a GARCH(1,1), with the correct unconditional variance, but with differing persistence. The setup is such that the ratio of the coefficient on r_{t-1}^2 to the coefficient on h_{t-1} is always equal to 10/85. We take these parameters as fixed, and do not consider the forecaster's problem of estimating the model parameters from the data. The null corresponds to the case that $k = 0.95$.

$$\begin{aligned} Model \ : \ h_t &= (1-k) + \frac{0.85}{0.95} k \times h_{t-1} + \frac{0.10}{0.95} k \times r_{t-1}^2 \\ k &= \{0.80, 0.81, \ldots, 0.99, 1\} \end{aligned} \tag{27}$$

We studied varying degrees of persistence (k) and 4 different sample sizes (T) designed to reflect realistic evaluation samples, ranging from 100 to 1000, each with 10,000 replications. The power curves are presented in Fig. 1, and the finite-sample size properties are summarised in Table 1.

Three main conclusions can be drawn from our small Monte Carlo study: Firstly, MZ1 tests are generally more powerful than the MZ2 tests, particu-

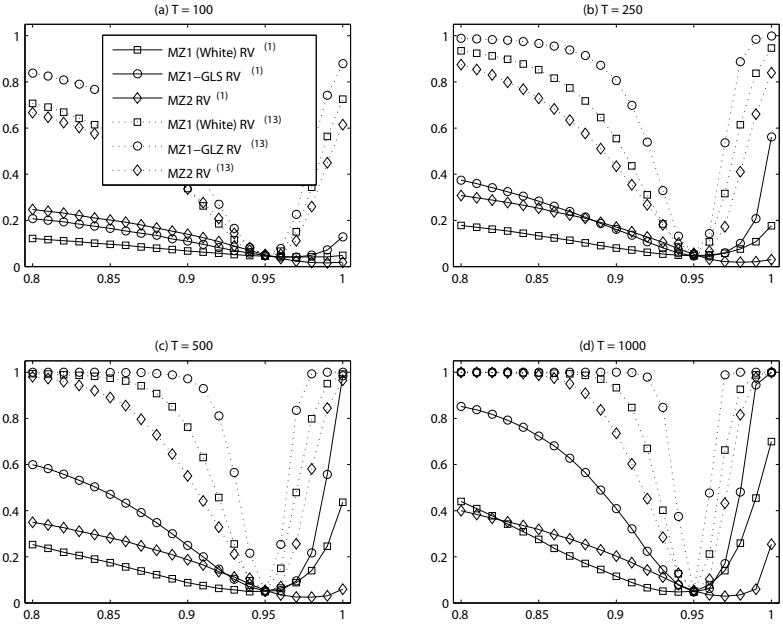

Fig. 1 The four panels of this figure contain the size-adjusted power for MZ1 and MZ2 tests applied to volatility forecasts. Two proxies were used for the latent volatility: daily squared returns ($RV^{(1)}$, *solid line*) and a 13-sample realised variance estimator ($RV^{(13)}$, *dotted line*). MZ1 tests regress the volatility proxy on the forecast using either OLS, indicated by □, or GLS, indicated by ○, and MZ2 tests regress the standardised volatility proxy on a lag of the standardised volatility proxy, indicated by ◇.

larly for larger sample sizes. Secondly, of the MZ1 tests, MZ1-GLS has better power, often substantially, than the standard feasible MZ test. The difference between MZ1 and MZ1-GLS was particularly striking when using the daily squared return as a proxy. Furthermore, it has good size properties even in small samples. That MZ1-GLS has better size properties than tests relying on robust standard errors is not too surprising, given that robust standard error estimators are known to often perform poorly in finite samples. Finally, the use of high-frequency data provides substantial gains in power: using just 13 intra-daily observations (corresponding to 30-minute returns for a stock traded on the NYSE) yields a marked improvement in power over a proxy based on daily returns. Using only $T = 100$ observations based on realised volatility with $m = 13$ produced similar power to using $T = 1000$ observations and squared daily returns ($m = 1$).

The size and power of the tests of covariance forecasts were examined in a similar study. Using the natural analogue of the GARCH(1,1), the DGP was

specified as a bivariate scalar diagonal *vech*,

$$\begin{aligned} DGP \;:\; \mathbf{r}_t &= \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\varepsilon}_t,\; t=1,2,\ldots,T \\ \text{where}\quad E_{t-1}\left[\boldsymbol{\varepsilon}_t\right] &= \mathbf{0} \text{ and } \; E_{t-1}\left[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right] = \mathbf{I}_k \\ \boldsymbol{\Sigma}_t &= 0.05\bar{\boldsymbol{\Sigma}} + 0.85\boldsymbol{\Sigma}_{t-1} + 0.10\mathbf{r}_{t-1}\mathbf{r}_{t-1}' \\ T &= \{100, 250, 500, 1000\} \end{aligned} \tag{28}$$

where $\bar{\boldsymbol{\Sigma}}$ is a bivariate matrix with unit diagonals and off-diagonal values of 0.3 (see Bollerslev et al. (1988)). The standardised innovations were specified to allow DGP-consistent realised covariances to be computed,

$$\begin{aligned} \boldsymbol{\varepsilon}_t &= \sum_{m=1}^{78} \boldsymbol{\xi}_{mt} \\ \boldsymbol{\xi}_{mt} &\overset{\text{IID}}{\sim} N\left(0, 1/78\mathbf{I}_2\right). \end{aligned} \tag{29}$$

The forecasting model was also a scalar diagonal *vech* parameterised in the spirit of (23),

$$\begin{aligned} \mathbf{H}_t &= (1-k)\,\bar{\boldsymbol{\Sigma}} + \frac{0.85}{0.95}k\mathbf{H}_{t-1} + \frac{0.10}{0.95}k\mathbf{r}_{t-1}\mathbf{r}_{t-1}' \\ k &= \{0.80, 0.81, \ldots, 0.99, 1\}. \end{aligned}$$

When $k = 0.95$ the model corresponds to the DGP. Daily cross-products and 13- and 78-sample realised covariance, $RC^{(13)}$ and $RC^{(78)}$, respectively, were used to proxy for the latent covariance and 10,000 replications were conducted.

In addition to the MZ1, MZ1-GLS and MZ2 specification studied in the volatility evaluation Monte Carlo, the approximate GLS specification in (21) was included in the study and is indicated by MZ1-Approx. GLS. Table 2 contains summary information about the size and Fig. 2 contains size-adjusted power curves for the specifications examined. The results are in line with those of the volatility tests: the tests have better size when GLS is used, particularly when the use of robust standard errors can be avoided, and the size is improved by using a more precise proxy. The power curves also show that MZ2 is less powerful than MZ1.

Combining the results of these two simulations, two main conclusions emerge. Firstly, the gains from using intra-daily data to construct a realised volatility proxy are large, even when only using a few intra-daily samples. Tests based solely on daily data are often oversized and have low power, even for $T = 1000$. The fact that substantial gains from intra-daily data may be obtained even when using just 30-minute returns is a positive result, given that prices sampled at this frequency are generally believed to be free from microstructure noise and non-synchronous trading problems; something that

Size Comparison of MZ-type Tests for Covariance Forecasts						
	T = 100			T = 250		
	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
MZ1 (White)	0.18	0.09	0.08	0.13	0.07	0.07
MZ1-GLS	0.10	0.06	0.06	0.07	0.05	0.05
MZ1-GLS (White)	0.12	0.07	0.07	0.08	0.06	0.06
MZ1-Approx GLS (White)	0.13	0.07	0.07	0.09	0.06	0.06
MZ2	0.05	0.05	0.05	0.05	0.05	0.05
MZ2 (White)	0.06	0.08	0.08	0.07	0.06	0.06
	T = 500			T = 1000		
	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
MZ1 (White)	0.11	0.06	0.06	0.09	0.06	0.06
MZ1-GLS	0.06	0.05	0.06	0.06	0.05	0.05
MZ1-GLS (White)	0.07	0.05	0.06	0.06	0.05	0.05
MZ1-Approx GLS (White)	0.07	0.05	0.06	0.06	0.05	0.05
MZ2	0.05	0.05	0.05	0.05	0.05	0.05
MZ2 (White)	0.05	0.05	0.05	0.05	0.05	0.05

Table 2 Rejection frequency of test statistics using a nominal size of 5% based on asymptotic critical values. All test statistics were computed using Wald tests using either White's heteroskedasticity-consistent covariance estimator or the standard OLS variance covariance formula as indicated. Three proxies were used for the latent covariance matrix: daily outer-products of returns ($RV^{(1)}$), a 13-sample realised covariance estimator ($RV^{(13)}$) and a 78-sample realised covariance estimator ($RV^{(78)}$).

is not true for prices sampled a 5-minute intervals (see, e.g. Hansen and Lunde (2006b), Griffin and Oomen (2006), Sheppard (2006)).

The second main finding is that the use of a GLS estimator produces substantial improvements in finite-sample size and distinct increases in power. For example, when using daily squared returns as the volatility proxy, the use of the a GLS regression is approximately as powerful as a standard MZ regression with twice as many observations.

3 Direct Comparison of Volatility Forecasts

This section reviews methods for comparing two or more competing volatility forecasts using a volatility proxy.

Direct comparisons of competing volatility forecasts can be done in a number of ways. One popular approach for univariate volatility forecasts is to compare forecasts using the R^2 from MZ regressions of a proxy on the forecast, see Andersen and Bollerslev (1998), Andersen et al. (2005). If the forecasts

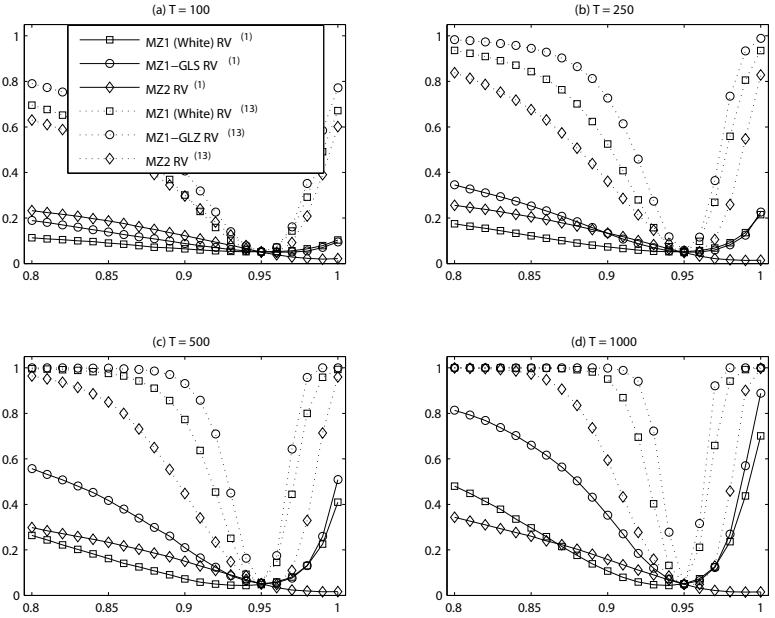

Fig. 2 The four panels of this figure contain the size-adjusted power for MZ1 and MZ2 tests applied to covariance forecasts. Two proxies were used for the latent covariance matrix: daily outer-products of returns ($RV^{(1)}$, *solid line*) and a 13-sample realised covariance estimator ($RV^{(13)}$, *dotted line*). MZ1 tests regress the volatility proxy on the forecast using either OLS, indicated by □, or GLS, indicated by ○, and MZ2 tests regress the standardised volatility proxy on a lag of the standardised volatility proxy, indicated by ◇.

are unbiased, this is equivalent to ranking the forecasts on the basis of their mean square error (MSE). The significance of any difference in MSE, or R^2, can be tested via a Diebold-Mariano and West (henceforth DMW) test (see Diebold and Mariano (1995) and West, (1996, 2006)).[11]

3.1 Pair–wise comparison of volatility forecasts

The DMW test can be used to compare two forecasts using general loss functions, including those other than MSE. Consider the general case, with

[11] West (1996) explicitly considers forecasts that are based on estimated parameters, whereas the null of equal predictive accuracy is based on population parameters. Diebold and Mariano (1995), on the other hand, take the forecasts as given and do not allow for estimation error. In this chapter we also take the forecasts as given, and so these two approaches coincide.

the loss function defined over the true conditional covariance, $\mathbf{\Sigma}_t$, or a proxy, $\widehat{\mathbf{\Sigma}}_t$, and the covariance forecast, $\mathbf{H}_t$, $L : \mathbb{M}_+^K \times \mathcal{H}^K \rightarrow \mathbb{R}_+$. The DMW tests the null of equal predictive accuracy against composite alternatives that indicate which forecast performs better:

$$\begin{aligned} & H_0 : E\left[L(\mathbf{\Sigma}_t, \mathbf{H}_t^A)\right] = E\left[L(\mathbf{\Sigma}_t, \mathbf{H}_t^B)\right] \\ \text{vs.} \quad & H_1 : E\left[L(\mathbf{\Sigma}_t, \mathbf{H}_t^A)\right] > E\left[L(\mathbf{\Sigma}_t, \mathbf{H}_t^B)\right] \\ & H_2 : E\left[L(\mathbf{\Sigma}_t, \mathbf{H}_t^A)\right] < E\left[L(\mathbf{\Sigma}_t, \mathbf{H}_t^B)\right] \end{aligned} \tag{30}$$

Since $\mathbf{\Sigma}_t$ is unobservable, this test is implemented using a statistic computed on the difference in the losses measured via a volatility proxy:

$$d_t = L(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t^A) - L(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t^B) \tag{31}$$

The test is computed using a standard t-test:

$$DMW_T = \frac{\sqrt{T}\bar{d}_T}{\sqrt{\widehat{avar}\left[\sqrt{T}\bar{d}_T\right]}} \tag{32}$$

$$\text{where } \bar{d}_T \equiv \frac{1}{T}\sum_{t=1}^{T} d_t$$

and $\widehat{avar}\left[\sqrt{T}\bar{d}_T\right]$ is some consistent estimator of the asymptotic variance of the re-scaled average, $\sqrt{T}\bar{d}_T$, such as the Newey-West estimator (Newey and West (1987)). Under the null hypothesis the test statistic is asymptotically Normally distributed.

Giacomini and White (2006) recently proposed comparing forecasts based on their expected loss *conditional* on variables thought to be important for relative forecast performance. The null hypothesis of interest in such a comparison replaces the unconditional expectation in (30) with a conditional expectation. For example, in a volatility forecasting application, one could test whether one volatility forecast out-performs another more in times of high volatility than in times of low volatility, or during bull markets compared to bear markets. However, we have not seen such an application in the literature to date.

3.2 Comparison of many volatility forecasts

When interest lies in comparing many competing forecasts there are two main approaches: the "Reality Check", cf. White (2000), and a modified version with better power properties, cf. Hansen (2005), or the "Model Confidence

Set" (Hansen et al. (2005)). The Reality Check tests the null hypothesis that no forecast outperforms, according to some loss function, a given benchmark forecast. The null hypothesis in this test is

$$\begin{aligned} H_0 : E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^A)\right] &\leq \min_{i \in \{B,C,\ldots\}} E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^i)\right] \\ \text{vs.} \quad H_1 : E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^A)\right] &> \min_{i \in \{B,C,\ldots\}} E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^i)\right] \end{aligned}$$

Hansen and Lunde (2005) use this type of test to determine whether a GARCH(1,1) model was out-performed by any of over three hundred competing volatility forecasts, for IBM equity return volatility and for the volatility of the log-difference of the DM-USD exchange rate .

The Model Confidence Set (MCS) is useful when there is no benchmark forecast. The outcome of this approach is a subset of forecasts that are not distinguishable from the best forecast across the complete set of forecasts. Defining the set of all competing forecasts as $\mathcal{M} = \{A, B, C, \ldots\}$, the MCS tests the null that no forecast is distinguishable against an alternative that at least one of the forecasts has a higher expected loss,

$$\begin{aligned} H_0 : E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^i)\right] &= E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^j)\right] \text{ for all } i, j \in \mathcal{M} \\ \text{vs.} \quad H_1 : E\left[L(\boldsymbol{\Sigma}_i, \mathbf{H}_t^i)\right] &> E\left[L(\boldsymbol{\Sigma}_t, \mathbf{H}_t^j)\right] \text{ for some } i \in \mathcal{M}, \text{for all } j \in \mathcal{M} \setminus i. \end{aligned}$$

The MCS operates by iteratively deleting poorly performing forecasts to construct a set, $\mathcal{M}^*$, that contains the forecast producing the lowest expected loss with probability weakly greater than the level of the test (e.g. 0.05), with the property that the probability that this set contains a sub-optimal forecast asymptotes to zero with the sample size. The MCS resembles in many respects a confidence interval for a parameter.

3.3 'Robust' loss functions for forecast comparison

Common to all approaches for comparing volatility forecasts is the focus on expected loss using the true, latent, covariance. In practice, however, the actual quantity computed is the difference in expected losses evaluated at some volatility proxy. Patton (2006) defines a loss function as "robust" if it yields the same ranking of competing forecasts using an unbiased volatility proxy, $E_{t-1}[\widehat{\boldsymbol{\Sigma}}_t] = \boldsymbol{\Sigma}_t$ as would be obtained using the (unobservable) conditional variance. Patton's focus was on the evaluation of univariate volatility forecasts, but the extension of his definition of loss function "robustness" is clear:

Definition 1 A loss function, L, is "robust" if the ranking of any two (possibly imperfect) volatility forecasts, $\mathbf{H}_t^A$ and $\mathbf{H}_t^B$, by expected loss is the same

whether the ranking is done using the true conditional variance, $\mathbf{\Sigma}_t$, or some conditionally unbiased volatility proxy, $\widehat{\mathbf{\Sigma}}_t$:

$$\begin{aligned} & E\left[L\left(\mathbf{\Sigma}_t, \mathbf{H}_t^A\right)\right] \gtreqless E\left[L\left(\mathbf{\Sigma}_t, \mathbf{H}_t^B\right)\right] \\ \Leftrightarrow & E\left[L\left(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t^A\right)\right] \gtreqless E\left[L\left(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t^B\right)\right] \end{aligned} \tag{33}$$

For univariate volatility forecast comparison, Meddahi (2001) showed that the ranking of forecasts on the basis of the R^2 from a standard Mincer-Zarnowitz regression is robust to noise in $\widehat{\sigma}_t^2$. Hansen and Lunde (2006a) showed that the R^2 from a regression of $\log\left(\widehat{\sigma}_t^2\right)$ on a constant and $\log\left(h_t\right)$ is not robust to noise. Moreover, they showed a sufficient condition for a loss function to be robust is that $\partial^2 L\left(\widehat{\sigma}_t^2, h_t\right)/\partial\left(\widehat{\sigma}_t^2\right)^2$ does not depend on h_t. Patton (2006) generalised this result by providing necessary and sufficient conditions for a univariate loss function to be robust.

3.4 Problems arising from 'non–robust' loss functions

In this section we investigate the problems caused by the use of non-robust loss functions in univariate volatility forecast comparison, and the reduction in the magnitude of these problems achieved through the use of higher frequency data (such as realised volatility). Patton showed that if a loss function is robust, then the optimal forecast under L, defined as

$$h_t^* \equiv \arg\min_{h \in \mathcal{H}} \; E_{t-1}\left[L\left(\widehat{\sigma}_t^2, h\right)\right]$$

must be the conditional variance (see Patton (2006)). One measure of the degree of distortion caused by the use of a loss function in combination with a noisy volatility proxy is the degree of bias in the optimal forecast under that loss function. Patton (2006) analytically derived the bias caused by nine widely-used loss functions, see equations (34) to (42) below, when combined with the range or realised variance as the volatility proxy. Under a simple zero-drift Brownian motion assumption for the return process he found that the (multiplicative) bias ranged from 0.28 to 3 using daily squared returns as the volatility proxy, but shrank to 0.98 to 1.03 if a realised variance estimator based on 5-minute returns was available.

To investigate whether these dramatic reductions in bias when using volatility proxies based on high frequency data hold under more realistic assumptions on the data generating process (DGP), we conduct a small simulation study. We consider three data generating processes, using the same models and parameter values as the simulation study of Gonçalves and Meddahi (2005). The first model is a GARCH diffusion, as in Andersen and Bollerslev (1998):

$$\begin{aligned} \mathrm{d}\log P_t &= 0.0314\mathrm{d}t + \nu_t\left(-0.576\mathrm{d}W_{1t} + \sqrt{1-0.576^2}\mathrm{d}W_{2t}\right) \\ \mathrm{d}\nu_t^2 &= 0.035\left(0.636-\nu_t^2\right)\mathrm{d}t + 0.144\nu_t^2\mathrm{d}W_{1t} \end{aligned}$$

The second model is a log-normal diffusion, as in Andersen et al. (2002):

$$\begin{aligned} \mathrm{d}\log P_t &= 0.0314\mathrm{d}t + \nu_t\left(-0.576\mathrm{d}W_{1t} + \sqrt{1-0.576^2}\mathrm{d}W_{2t}\right) \\ \mathrm{d}\log\nu_t^2 &= -0.0136\left(0.8382+\log\nu_t^2\right)\mathrm{d}t + 0.1148\mathrm{d}W_{1t} \end{aligned}$$

The final model is the two-factor diffusion, see Chernov et al. (2003):

$$\begin{aligned} \mathrm{d}\log P_t &= 0.030\mathrm{d}t + \nu_t\left(-0.30\mathrm{d}W_{1t} - 0.30\mathrm{d}W_{2t}\right. \\ &\qquad \left. + \sqrt{1-0.3^2-0.3^2}\mathrm{d}W_{3t}\right) \\ \nu_t^2 &= \text{s-exp}\left\{-1.2 + 0.04\nu_{1t}^2 + 1.5\nu_{2t}^2\right\} \\ \mathrm{d}\nu_{1t}^2 &= -0.00137\nu_{1t}^2\mathrm{d}t + \mathrm{d}W_{1t} \\ \mathrm{d}\nu_{2t}^2 &= -1.386\nu_{2t}^2\mathrm{d}t + \left(1+0.25\nu_{2t}^2\right)\mathrm{d}W_{2t} \end{aligned}$$

$$\text{where} \quad \text{s-exp}\left\{x\right\} = \begin{cases} \exp\left\{x\right\}, & x \le x_0 \\ \exp\left\{x_0\right\}\sqrt{1-x_0+x^2/x_0}, & x > x_0 \end{cases}$$

In simulating from these processes we use a simple Euler discretization scheme, with the step size calibrated to one second (i.e., with 23,400 steps per simulated day). The loss functions we consider are:

$$MSE: L\left(\widehat{\sigma}_t^2, h_t\right) = \left(\widehat{\sigma}_t^2 - h_t\right)^2 \tag{34}$$

$$QLIKE: L\left(\widehat{\sigma}_t^2, h_t\right) = \log h_t + \frac{\widehat{\sigma}_t^2}{h_t} \tag{35}$$

$$MSE\text{-}LOG: L\left(\widehat{\sigma}_t^2, h_t\right) = \left(\log\widehat{\sigma}_t^2 - \log h_t\right)^2 \tag{36}$$

$$MSE\text{-}SD: L\left(\widehat{\sigma}_t^2, h\right) = \left(\widehat{\sigma}_t - \sqrt{h_t}\right)^2 \tag{37}$$

$$MSE\text{-}prop: L\left(\widehat{\sigma}_t^2, h_t\right) = \left(\frac{\widehat{\sigma}_t^2}{h_t} - 1\right)^2 \tag{38}$$

$$MAE: L\left(\widehat{\sigma}_t^2, h_t\right) = \left|\widehat{\sigma}^2 - h_t\right| \tag{39}$$

$$MAE\text{-}LOG: L\left(\widehat{\sigma}_t^2, h_t\right) = \left|\log\widehat{\sigma}_t^2 - \log h_t\right| \tag{40}$$

$$MAE\text{-}SD: L\left(\widehat{\sigma}_t^2, h_t\right) = \left|\widehat{\sigma}_t - \sqrt{h_t}\right| \tag{41}$$

$$MAE\text{-}prop: L\left(\widehat{\sigma}_t^2, h_t\right) = \left|\frac{\widehat{\sigma}^2}{h_t} - 1\right| \tag{42}$$

In the presence of stochastic volatility, as in the three DGPs above, the appropriate volatility concept changes from the conditional variance of daily returns to the expected *integrated variance*, see Andersen et al. (2006b):

$$E_{t-1}\left[IV_t\right] \equiv E_{t-1}\left[\int_{t-1}^{t} \nu_\tau^2 d\tau\right]$$

We consider three realised volatility proxies, based on $m = 1, 13$ and 78 intra-daily observations. We also consider the use of the adjusted squared range as a volatility proxy, which is defined as

$$RG_t^{*2} = \frac{1}{4\log 2} RG_t^2$$
$$\text{where} \quad RG_t \equiv \max_\tau \log P_\tau - \min_\tau \log P_\tau, \; t-1 < \tau \leq t$$

The adjustment factor $\left(\frac{1}{4}\log 2\right)$ is required so as to make the squared range unbiased for the daily volatility for a Brownian motion with no drift, cf. Parkinson (1980). Christensen and Podolskij (2006) show that this adjustment factor leads to an asymptotically unbiased proxy for general Brownian semi-martingales (as the number of intra-daily observations, m, goes to infinity).

We present the results of our simulation in Table 3, and for ease of comparison we also present the analytical results from Patton for the Brownian motion case (Patton (2006)). This table contains the (multiplicative) biases in the optimal forecasts under the loss functions listed above, for the various DGPs. An unbiased forecast will have coefficient of one. The MSE and QLIKE loss functions, as expected, did not generate bias for any volatility proxy. These loss functions are easily shown to be "robust", and so lead to zero bias as long as the volatility proxy is unbiased.

The first three panels of Table 3 reveal that allowing for stochastic volatility through a GARCH diffusion or a log-Normal diffusion does not substantially change the degree of bias in optimal forecasts under various loss function/volatility proxy combinations relative to the simple Brownian motion case. In fact, almost all of the differences occur only in the second decimal place. This suggests that the biases computed under the simplistic assumption of constant intra-daily volatility are a good approximation to those obtained under GARCH or log-Normal SV DGPs.

The situation is quite different when the two-factor SV model, see Chernov et al. (2003) is considered. This model was developed to replicate the jump-like features observed in some data without actually introducing a jump component into the model and can generate extreme observations and excess kurtosis. Patton found that excess kurtosis generally exacerbated any biases that were present under normally distributed returns, and this is reinforced by our simulation results: the two-factor diffusion generates biases ranging from 0.35 to 6.70 even when using a realised volatility estimator based on 5-minute returns (Patton (2006)). That is, the biases from 5-minute realised volatility under the two-factor diffusion are greater than the biases from using *daily* squared returns when returns are conditionally normally distributed.

Biases in optimal volatility forecasts								
Loss function	Range	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	Range	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
	Brownian Motion				GARCH-SV			
MSE	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
QLIKE	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
MSE-LOG	0.85	0.28	0.91	0.98	0.83	0.28	0.92	0.98
MSE-SD	0.92	0.56	0.96	0.99	0.91	0.63	0.96	0.99
MSE-prop	1.41	3.00	1.15	1.03	1.40	3.02	1.16	1.03
MAE	0.83	0.45	0.95	0.99	0.82	0.46	0.94	0.99
MAE-log	0.83	0.45	0.95	0.99	0.82	0.46	0.94	0.99
MAE-SD	0.83	0.45	0.95	0.99	0.82	0.46	0.94	0.99
MAE-prop	0.99	2.36	1.10	1.02	1.18	2.37	1.10	1.01
	Log-normal SV				Two-factor SV			
MSE	0.99	1.00	1.00	1.00	1.00	1.01	1.00	1.00
QLIKE	0.99	1.00	1.00	1.00	1.00	1.01	1.00	1.00
MSE-LOG	0.83	0.28	0.92	0.98	0.35	0.12	0.37	0.41
MSE-SD	0.91	0.63	0.96	0.99	0.57	0.40	0.58	0.62
MSE-prop	1.40	3.03	1.16	1.03	9.79	20.60	9.03	6.70
MAE	0.82	0.46	0.94	0.99	0.31	0.17	0.32	0.35
MAE-LOG	0.82	0.46	0.94	0.99	0.31	0.17	0.32	0.35
MAE-SD	0.82	0.46	0.94	0.99	0.31	0.17	0.32	0.35
MAE-prop	1.18	2.37	1.10	1.02	3.47	6.60	3.33	2.98

Table 3 Bias, as a percentage of the underlying integrated variance, when using robust and non-robust loss functions. A coefficient of 1 indicates the loss function does not generate a biased optimal forecast. The results in the Brownian motion case are analytical while the results in the three other cases are the result of a simulation.

This simulation study shows that the use of volatility proxies with less noise (such as those based on higher-frequency data) ameliorates but does not eliminate the biases caused by the use of non-robust loss functions. The remaining biases can still be large depending on the form of the data generating process. This suggests that using robust loss functions in forecast comparison tests is important, even when higher quality volatility proxies are available.

3.5 Choosing a "robust" loss function

Patton (2006) derives a parametric class of univariate loss functions that are both homogeneous and robust to the use of a noisy volatility proxy. Homogeneity is a desirable property of a loss function that ensures the choice of normalisation (for example using raw returns versus 100 times returns) does not affect the optimal forecast beyond a known scale factor. The class can be described as[12]

$$L(\widehat{\sigma}_t^2, h_t; b) = \begin{cases} \frac{1}{(b+1)(b+2)}(\widehat{\sigma}_t^{2(b+2)} - h_t^{b+2}) \\ \qquad -\frac{1}{b+1} h_t^{b+1}(\widehat{\sigma}_t^2 - h_t) & , b \neq -1, -2 \\ h_t - \widehat{\sigma}_t^2 + \widehat{\sigma}_t^2 \log \frac{\widehat{\sigma}_t^2}{h_t} & , b = -1 \\ \frac{\widehat{\sigma}_t^2}{h_t} - \log \frac{\widehat{\sigma}_t^2}{h_t} - 1 & , b = -2 \end{cases} \tag{43}$$

The ability of these loss functions to distinguish between competing volatility forecasts in a DMW test may vary substantially with the choice of "shape" parameter, b. To provide some insight into this problem, we conduct a small simulation study of the size and power of a DMW test of the null that two competing forecasts have equal predictive power. The models we consider are based on those used in the Mincer-Zarnowitz Monte Carlo study in Section 2.5. The returns were generated according to (23). The first forecast, h_t^A, is set equal to the conditional variance of the process multiplied by an *iid* error term, z_t^A, distributed as a standardised χ^2 random variable: $\nu z_t^A \overset{IID}{\sim} \chi_\nu^2$. This error term can be thought of as representing (in a simple and computationally convenient way) estimation error or model mis-specification in the volatility forecast, for example. The second volatility forecast, h_t^B, is generated by (27), and is also multiplied by an *iid* error term, z_t^B, independent of the first, distributed as a standardised χ^2 random variable: $\nu z_t^B \overset{IID}{\sim} \chi_\nu^2$. We set the degree of freedom parameter, ν, to 500, which implies that the *iid* error terms have unit mean and standard deviations of 0.06. Although neither of these forecasts is perfect, h_t^A is weakly preferred to h_t^B. The point where h_t^B is also "correctly specified" ($k = 0.95$) corresponds to the case when the two forecasts are equally accurate.

Using the two forecasts, the loss was computed using the "robust" loss function with shape parameter $b \in \{-5, -3, -2, -1, 0, 2\}$, and a DMW test

[12] All loss functions have been normalised to 0 when $h_t = \widehat{\sigma}_t^2$. In applications where $\Pr(\widehat{\sigma}_t^2 = 0) > 0$ the normalised loss is not always well-defined. The normalisation terms can be removed without affecting any results with respect to robustness. Homogeneity is also preserved, effectively, as the removal of these normalising terms means that re-scaling the data adds a constant to the loss, which does not affect the optimum and drops out in forecast comparisons.

statistic was computed using the difference in the losses

$$d_t(b) = L(\widehat{\sigma}_t^2, h_t^A; b) - L(\widehat{\sigma}_t^2, h_t^B; b) \tag{44}$$

$$DMW_T(b) = \frac{\sqrt{T}\bar{d}_T(b)}{\sqrt{\widehat{avar}\left[\sqrt{T}\bar{d}_T(b)\right]}} \tag{45}$$

The asymptotic variance of the average was computed using a Newey-West variance estimator with the number of lags set to $\lceil T^{1/3} \rceil$.

Three proxies were used to compare the performance of the two forecasts: the daily squared return, $RV^{(1)}$, a 30-minute realised variance $RV^{(13)}$, and a 5-minute realised variance $RV^{(78)}$. The finite-sample size of the test is reported in Table 4. Overall the size of the test is good, however for larger values of b the test is undersized. This may be due to a lack of moments in the d_t series, which results in a non-standard distribution of the DMW test statistic. This non-standard behaviour was foreshadowed in Patton, who showed that $8 + \delta$ moments of returns are needed for a DMW test using squared returns as the volatility proxy, see Patton (2006). Larger values of b require even more moments of returns. In many realistic cases, including the model use in the simulation study, returns will not be sufficiently well behaved for tests based loss functions $b \geq 0$ to be reliable.

The power of the test was studied using a model with mis-specified dynamics where the alternative forecasts were generated from either over- or under-persistent models. Figure 3 contains four views into the power of the test using $T = 100$ and $T = 250$ using daily squared returns and 13-sample realised volatility. Panels (c) and (d) confirm that improved proxies for volatility have a distinct effect on the ability to discern superior forecast performance.

Three of the four panels show that the QLIKE loss function ($b = -2$) has the highest power. Using either daily returns or 30-minute realised volatility, power drops off markedly when using large b loss functions. Even when using a precise proxy and a long sample there is a distinct decrease in power for the loss functions with b furthest from -2.

Figure 4 contains the average size-adjusted power (averaging the power curves in Fig. 3 across all $k \in \{0.80, 0.81, \ldots, 1.00\}$) as a function of b using daily returns squared and 30-minute realised variance. Both power plots exhibit peaks near -2 in all cases except the smallest sample size test statistic computed from daily returns. These results all point to QLIKE as the preferred choice amongst the loss functions that are both homogeneous and robust to noise in the proxy[13].

[13] It is worth noting that the results presented for $b = -2$ are likely a close-to-ideal case. Returns in practice are decidedly non-normal exhibiting fat-tails, skewness and/or jumps. The increased propensity for actual asset returns to produce large observations should produce even worse performance of loss functions with $b \geq 0$.

Finite-Sample Size of DMW Tests						
	T = 100			T = 250		
	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
b=-5	0.05	0.06	0.05	0.05	0.05	0.05
b=-3	0.06	0.07	0.07	0.06	0.06	0.06
b=-2 (QLIKE)	0.06	0.07	0.07	0.06	0.06	0.06
b=-1	0.06	0.07	0.07	0.05	0.06	0.06
b=0 (MSE)	0.05	0.06	0.05	0.04	0.04	0.05
b=2	0.03	0.03	0.03	0.02	0.02	0.02
	T = 500			T = 1000		
	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$	$RV^{(1)}$	$RV^{(13)}$	$RV^{(78)}$
b=-5	0.05	0.05	0.05	0.05	0.05	0.05
b=-3	0.05	0.05	0.06	0.05	0.05	0.05
b=-2 (QLIKE)	0.05	0.05	0.06	0.05	0.05	0.05
b=-1	0.05	0.05	0.05	0.05	0.05	0.05
b=0 (MSE)	0.04	0.04	0.04	0.04	0.04	0.04
b=2	0.02	0.02	0.02	0.02	0.02	0.02

Table 4 Finite-sample size of the DMW tests using different volatility proxies and sample sizes. Data was simulated according to a GARCH(1,1) and the series of forecasts, $\{h_t\}$, was produced. The size was computed by comparing the performance of a DMW test comparing the loss of $h_t^A = z_t^A h_t$ and $h_t^B = z_t^B h_t$ where z_t^A and z_t^B were independent standardised χ^2_ν random variables with $\nu = 500$.

3.6 Robust loss functions for multivariate volatility comparison

The homogeneous, robust loss functions of Patton (2006) all have a first-order condition of the form:

$$\frac{\partial L\left(\widehat{\sigma}_t^2, h_t; b\right)}{\partial h_t} = -h_t^b(\widehat{\sigma}_t^2 - h_t)\ . \tag{46}$$

From this first-order condition it is simple to see that when $h_t = \sigma_t^2\ a.s.$ and $h_t > 0$, the expected score is zero and the second derivative is positive. As a result, the true conditional variance is the solution to the expected loss minimisation problem. Extending this analysis to the evaluation of covariance forecasts is straightforward: The direct analogue of (46) for conditional covariance forecasts is

$$\frac{\partial L\left(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t; b\right)}{\partial \mathbf{H}_t} = -\mathbf{H}_t^b(\widehat{\mathbf{\Sigma}}_t - \mathbf{H}_t)\ . \tag{47}$$

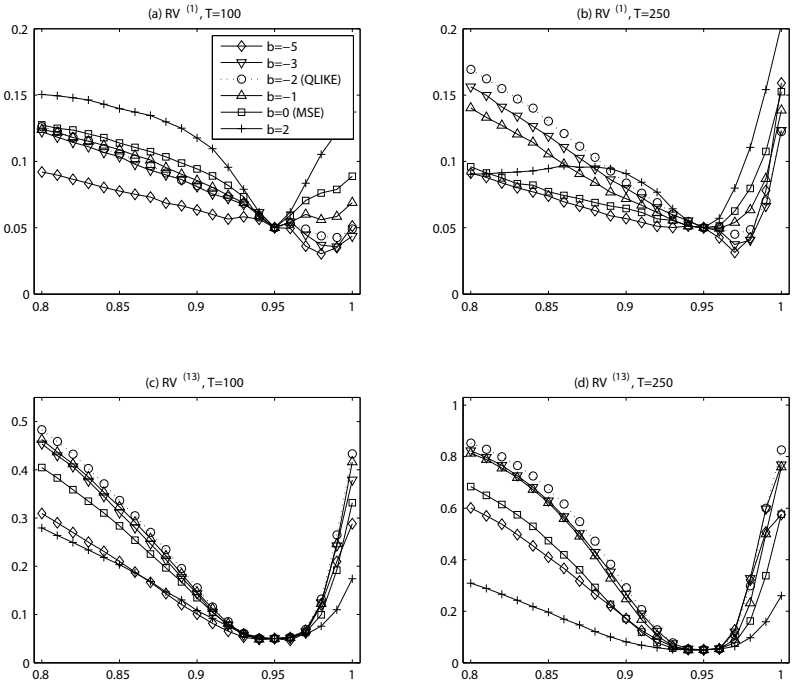

Fig. 3 Plots of size-adjusted power for DMW tests using two different proxies for latent volatility, daily returns squared ($RV^{(1)}$) and 30-minute realised volatility ($RV^{(13)}$). All loss functions are members of the "robust" loss function family, equation (43). Note: The scale of the y-axis changes in each panel.

While this is not the FOC of any standard expression, it does provide guidance to one class of FOCs, namely

$$\frac{\partial L\left(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t; b\right)}{\partial \mathbf{H}_t} = -\mathbf{C}_1\left(\mathbf{H}_t\right)\left(\widehat{\mathbf{\Sigma}}_t - \mathbf{H}_t\right)\mathbf{C}_2\left(\mathbf{H}_t\right) \ . \tag{48}$$

where $\mathbf{C}_1\left(\mathbf{H}_t\right)$ and $\mathbf{C}_2\left(\mathbf{H}_t\right)$ are positive definite matrix valued functions, $\mathbf{C}: \mathbb{M}_{++}^K \rightarrow \mathbb{M}_{++}^K$ for $i = 1, 2$, that do not depend on $\widehat{\mathbf{\Sigma}}_t$. Using the "vec" function to express this FOC as a column vector, it can be equivalently written

$$\frac{\partial L(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t; b)}{\partial \mathrm{vec}(\mathbf{H}_t)} = -(\mathbf{C}_2\left(\mathbf{H}_t\right) \otimes \mathbf{C}_1\left(\mathbf{H}_t\right))\mathrm{vec}(\widehat{\mathbf{\Sigma}}_t - \mathbf{H}_t) \ . \tag{49}$$

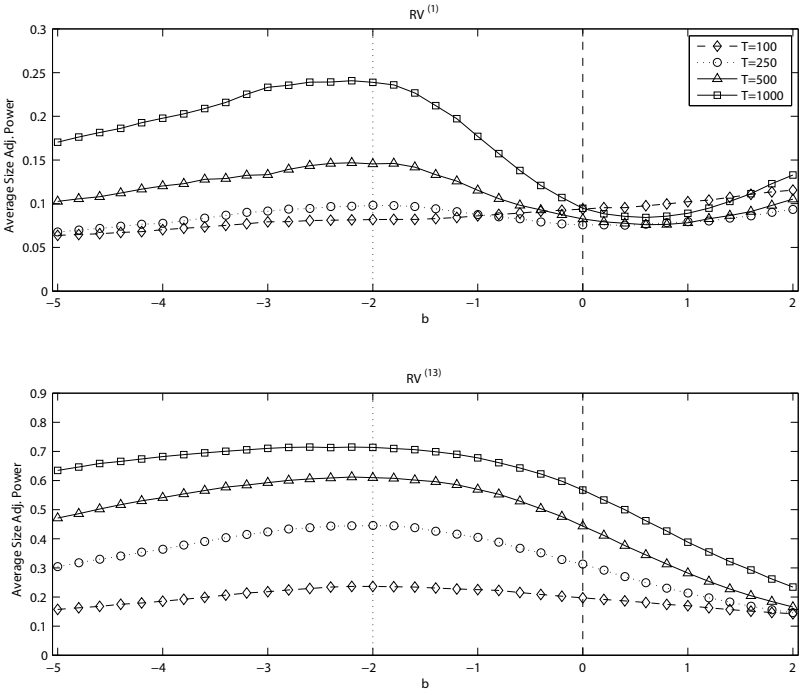

Fig. 4 Average size-adjusted power curves for DMW tests as a function of b. DMW tests were computed using daily squared returns ($RV^{(1)}$, top panel) or 13-sample realised volatility ($RV^{(13)}$, bottom panel) using the "robust" loss function family in equation (43). Note: The scale of the y-axis changes in each panel.

It is simple to verify that this FOC will be zero as long as $E_{t-1}[\widehat{\mathbf{\Sigma}}_t] = \mathbf{\Sigma}_t = \mathbf{H}_t$ *a.s.* The second order derivative with respect to $\mathbf{H}_t$ is $\mathbf{C}_2(\mathbf{H}_t) \otimes \mathbf{C}_1(\mathbf{H}_t)$ since $\partial \text{vec}\mathbf{H}_t / \partial \mathbf{H}_t = \mathbf{I}_{K^2}$ and is positive semi-definite by construction. The natural analogue to the class of robust, homogeneous loss functions introduced in Patton (2006) is thus

$$L(\widehat{\mathbf{\Sigma}}_t, \mathbf{H}_t; b) = \begin{cases} \frac{2}{(b+2)} \text{tr}(\mathbf{\Sigma}_t^{(b+2)} - \mathbf{H}_t^{b+2}) - \text{tr}(\mathbf{H}^{b+1}(\mathbf{\Sigma}_t - \mathbf{H}_t)) & b \neq -1, -2 \\ \text{tr}(\mathbf{H}_t^{-1}\mathbf{\Sigma}_t) - \log|\mathbf{H}_t^{-1}\mathbf{\Sigma}_t| - K & b = -2 \end{cases} \tag{50}$$

As in the univariate case, this class nests both the multivariate QLIKE and multivariate MSE classes. To verify that this class satisfies the condition of (47), consider the first-order conditions when b is an integer[14]

$$\frac{\partial L(\widehat{\boldsymbol{\Sigma}}_t, \mathbf{H}_t; b)}{\partial \mathbf{H}_t} = \begin{cases} -\sum_{j=0}^{b} \mathbf{H}^j \boldsymbol{\Sigma}_t \mathbf{H}^{b-j} - \mathbf{H}_t^{b+1} & b \in \mathbb{Z}_+ \\ -\mathbf{H}_t^{-1} \boldsymbol{\Sigma}_t \mathbf{H}_t^{-1} + \mathbf{H}_t^{-1} & b = -2 \end{cases} \tag{51}$$

and, transforming to column vectors,

$$\frac{\partial L(\widehat{\boldsymbol{\Sigma}}_t, \mathbf{H}_t; b)}{\partial \operatorname{vec}(\mathbf{H}_t)} = -\begin{cases} -\sum_{j=0}^{b} \left(\mathbf{H}^{b-j} \otimes \mathbf{H}^j\right) \operatorname{vec}(\widehat{\boldsymbol{\Sigma}}_t - \mathbf{H}_t) & b \in \mathbb{Z}_+ \\ -\left(\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}\right) \operatorname{vec}(\widehat{\boldsymbol{\Sigma}}_t - \mathbf{H}_t) & b = -2 \end{cases} \tag{52}$$

It should be noted that, unlike the univariate case, this class of loss functions does not encompass all that are robust and homogeneous. The expanded possibilities arise naturally because there are many functions $\mathbf{C}$ that can be used to weight the forecast errors, $\operatorname{vec}(\widehat{\boldsymbol{\Sigma}}_t - \mathbf{H}_t)$.

3.7 Direct comparison via encompassing tests

Encompassing tests are an alternative to DMW tests for comparing the performance of two or more forecasts. Rather than compare the loss of using one forecast to the loss of using the other, encompassing tests examine whether some function, generally affine, of the forecasts produces a smaller loss than a single forecast. Restricting attention to only two forecasts, the null and alternative tests are

$$H_0^A : E[L(\widehat{\sigma}_t^2, h_t^A)] = E[L(\widehat{\sigma}_t^2, f(h_t^A, h_t^B; \boldsymbol{\theta}))] \tag{53}$$
$$\text{vs. } H_1^A : E[L(\widehat{\sigma}_t^2, h_t^A)] > E[L(\widehat{\sigma}_t^2, f(h_t^A, h_t^B; \boldsymbol{\theta}))] \tag{54}$$

$$H_0^B : E[L(\widehat{\sigma}_t^2, h_t^B)] = E[L(\widehat{\sigma}_t^2, f(h_t^A, h_t^B; \boldsymbol{\theta}))] \tag{55}$$
$$\text{vs. } H_1^B : E[L(\widehat{\sigma}_t^2, h_t^B)] > E[L(\widehat{\sigma}_t^2, f(h_t^A, h_t^B; \boldsymbol{\theta}))] \tag{56}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters. The "forecast combination function", $f(h_t^A, h_t^B; \boldsymbol{\theta})$ is typically specified as a linear combination of the forecasts, thus $f(h_t^A, h_t^B; \boldsymbol{\theta}) = \beta_1 h_t^A + \beta_2 h_t^B$. When using the MSE loss function,

[14] For clarity of exposition, we only show the FOC when b is an integer greater than 0. The cases with non-integer b involve eigenvalue decompositions and are too lengthy to present here, although the loss function analysis goes through unmodified for any $b \neq -1$.

the encompassing test reduces to a standard augmented MZ regressions (see (9)) where the $\mathcal{F}_t$ measurable "instruments" are the competing forecasts,

$$\widehat{\sigma}_t^2 = \beta_1 h_t^A + \beta_2 h_t^B + u_t \tag{57}$$

and testing

$$H_0^A : \beta_1 = 1 \cap \beta_2 = 0$$
$$H_1^A : \beta_1 \neq 1 \cup \beta_2 \neq 0$$

To test whether forecast B encompasses forecast A we test:

$$H_0^B : \beta_1 = 0 \cap \beta_2 = 1$$
$$H_1^B : \beta_1 \neq 0 \cup \beta_2 \neq 1$$

Like the augmented MZ, the performance of this specification will generally be adversely affected by heteroskedastic errors. A GLS version can be specified

$$\frac{\widehat{\sigma}_t^2}{\widetilde{h}_t} = \beta_1 \frac{h_t^A}{\widetilde{h}_t} + \beta_2 \frac{h_t^B}{\widetilde{h}_t} + \widetilde{u}_t \tag{58}$$

where $\widetilde{h}_t = h_t^A$ to test H_0^A, and $\widetilde{h}_t = h_t^B$ to test H_0^B. If there is no natural null hypothesis, there are two alternative choices for the weights, $\widetilde{h}_t$: The first employs an average of h_t^A and h_t^B, either geometric $\widetilde{h}_t = \sqrt{h_t^A h_t^B}$ or arithmetic $\widetilde{h}_t = \frac{1}{2}h_t^A + \frac{1}{2}h_t^B$. The second uses a two-step feasible GLS (FGLS) where (57) is initially estimated and then (58) is estimated using $\widetilde{h}_t = \widehat{\beta}_1 h_t^A + \widehat{\beta}_2 h_t^B$, although care is needed to ensure the fitted volatilities are positive[15].

An alternative specification for $f(h_t^A, h_t^B; \boldsymbol{\theta})$, one that avoids any problems of negative fit volatilities, uses a geometric-average inspired form:

$$f(h_t^A, h_t^B; \boldsymbol{\theta}) = \exp(\beta_1 \ln h_t^A + \beta_2 \ln h_t^B) \tag{59}$$

The null and alternative hypotheses are identical, although non-linear least squares estimation is needed to implement this form of an encompassing test.

Extending encompassing tests to covariance forecasts is straightforward. The only remaining choice is the combination function. The natural candidate is a linear function:

$$f(\mathbf{H}_t^A, \mathbf{H}_t^B; \boldsymbol{\theta}) = \beta_1 \mathbf{H}_t^A + \beta_2 \mathbf{H}_t^B \ . \tag{60}$$

[15] While it does not affect the asymptotic distribution, there may be finite sample gains to using improved first-stage estimates from a standard GLS assuming pre-specified weights, such as $\widetilde{h}_t = h_t^A$ or $\widetilde{h}_t = h_t^B$.

Using a linear combination function, encompassing can be tested using an augmented MZ version of the pooled panel regression (see 16). An alternative choice for the combination covariance would be to use a geometric average,

$$f(\mathbf{H}_t^A, \mathbf{H}_t^B; \boldsymbol{\theta}) = \text{expm}\left(\beta_1 \text{logm}\mathbf{H}_t^A + \beta_2 \text{logm}\mathbf{H}_t^B\right) \tag{61}$$

where "expm" and "logm" are matrix exponentiation and logarithm respectively, see Magnus and Neudecker (2002). This alternative specification removes any restrictions on the estimated parameters while ensuring that the combination is strictly positive definite.

4 Indirect Evaluation of Volatility Forecasts

While statistical evaluation of forecasts is useful for ranking volatility forecasts, most forecasts are designed to aid in economic applications, see Andersen, Bollerslev, Christoffersen and Diebold (Andersen et al. (2006a)) for a recent survey of volatility and correlation forecasting. Economic evaluation of volatility forecasts is an important metric for assessing the performance of models.

Volatility and covariance forecasts are fundamental inputs into many decisions in financial economics: mean-variance portfolio optimisation, hedging, risk measurement, option pricing and utility maximisation all rely on forecast variances and covariance as inputs. While volatility forecasts have been evaluated using all of these applications, mean-variance portfolio choice and hedging are unique in that the correct conditional volatility or covariance, $\boldsymbol{\Sigma}_t = \mathbf{H}_t$ *a.s.*, will lead to improved performance without strong auxiliary assumptions. In general, economic evaluation of volatility and correlation forecasts relies in an important way on other assumptions, such as the utility function of the hypothetical investor (in portfolio choice or hedging applications), the density of the standardised returns (as in Value-at-Risk and Expected Shortfall forecasting, density forecasting, portfolio choice applications with non-quadratic utility), or the derivative pricing model (in option, and other derivative securities, pricing applications). Thus these approaches are "non-robust", in the sense described in the Introduction, however with strong economic motivations they can yield valuable information on competing volatility and correlation forecasts.

Economic evaluation of covariance forecasts has traditionally focused on the out-of-sample performance of portfolios formed using the forecast as an input in a mean-variance framework. Noting the sensitivity of portfolio weights to small changes in the conditional mean, many have questioned the wisdom of conditional mean-variance optimisation. However, recent work by Engle and Colacito (2006) (EC, henceforth) has clarified the role that the uncertain mean plays in comparing the performance of two or more forecasts. The

EC framework establishes that the correct covariance will produce a smaller expected portfolio variance for *any* assumed non-zero vector of means. As a result, evidence of superiority based on ex-post measurements of portfolio variance can be attributed to the covariance forecast.

4.1 Portfolio optimisation

Portfolio optimisation is a natural application for covariance forecasts. To avoid specifying or estimating asset mean returns, many authors have focussed on the problem of finding the global minimum variance portfolio (GMVP). It is computed as the solution to

$$\min_{\mathbf{w}_t} \mathbf{w}_t' \boldsymbol{\Sigma}_t \mathbf{w}_t \qquad \text{subject to } \mathbf{w}_t' \boldsymbol{\iota} = 1$$

It is simple to show that if the portfolio weights, $\mathbf{w}_t$, are constructed from the true covariance $\mathbf{H}_t = \boldsymbol{\Sigma}_t$ *a.s.*, then the variance of a portfolio computed using the GMVP from any other forecast, $\widetilde{\mathbf{w}}_t$, must be larger:

$$\begin{aligned} V[\widetilde{\mathbf{w}}_t' \boldsymbol{\Sigma}_t \widetilde{\mathbf{w}}_t] =& (\mathbf{w_t} + \mathbf{c_t})' \boldsymbol{\Sigma}_t (\mathbf{w}_t + \mathbf{c_t}) \\ &= (\frac{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} + \mathbf{c}_t') \boldsymbol{\Sigma}_t (\frac{\boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} + \mathbf{c}_t) \\ &= \frac{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} \boldsymbol{\Sigma}_t \frac{\boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} + \frac{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} \boldsymbol{\Sigma}_t \mathbf{c}_t + \mathbf{c}_t' \boldsymbol{\Sigma}_t \frac{\boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} + \mathbf{c}_t' \boldsymbol{\Sigma}_t \mathbf{c}_t \\ &= \frac{1}{\boldsymbol{\iota}' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} + \mathbf{c}_t' \boldsymbol{\Sigma}_t \mathbf{c}_t > \frac{1}{\boldsymbol{\iota} \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\iota}} \end{aligned}$$

since $\boldsymbol{\iota}' \mathbf{c}_t = 0$ follows from $\boldsymbol{\iota}' \mathbf{w}_t = \boldsymbol{\iota}' \widetilde{\mathbf{w}_t} = 1$. Using this result, it is then possible to compare two competing covariance forecasts by comparing the volatility of the minimum variance portfolio constructed using each forecast. Let $\mathbf{w}^* (\mathbf{H}_t)$ denote the solution to (62) for covariance matrix $\mathbf{H}_t$, and define

$$d_t \equiv \mathbf{w}^* \left(\mathbf{H}_t^B\right)' \mathbf{r}_t \mathbf{r}_t' \mathbf{w}^* \left(\mathbf{H}_t^B\right) - \mathbf{w}^* \left(\mathbf{H}_t^A\right)' \mathbf{r}_t \mathbf{r}_t' \mathbf{w}^* \left(\mathbf{H}_t^A\right)$$

This then allows for a DMW forecast comparison test, as in (30): if the mean of d_t is significantly positive (negative) then forecast A (B) is the better forecast. If more than one competing forecast is being considered, the "reality check" or MCS can be used to test for superior performance (see Sect. 3.2).

The general mean-variance portfolio optimisation problem is

$$\min_{\mathbf{w}_t} \ \mathbf{w}_t' \boldsymbol{\Sigma}_t \mathbf{w}_t \qquad \text{subject to } \mathbf{w}_t' \boldsymbol{\mu}_t = \mu_0$$

where $\boldsymbol{\mu}_t = E_{t-1}[\mathbf{r}_t]$. If returns are expressed in excess of the risk free rate, optimal portfolio weights can be computed as

$$\mathbf{w}_t = \frac{\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}{\boldsymbol{\mu}_t\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}\mu_0$$

where the weight on the risk free asset is $1 - \mathbf{w}_t'\boldsymbol{\iota}$.[16] Recent work by Engle and Colacito (2006) has shown that it is possible to rank covariance forecasts using this more general mean-variance optimisation problem. These authors show that if expected returns are constant, $E_{t-1}[\mathbf{r}_t] = \boldsymbol{\mu}\ \forall\ t$, then for *any* assumed mean vector ($\boldsymbol{\mu}$) the true covariance will produce a portfolio with a variance less than or equal to that of any forecast where $\mathbf{H}_t \neq \boldsymbol{\Sigma}_t$. Engle & Colacito suggest using only positive values for $\boldsymbol{\mu}$, and, rather than testing using a single value, a quasi-Bayesian method can be used to integrate over a range of plausible values for $\boldsymbol{\mu}$.

Forecast evaluation tests, as opposed to comparison tests, are more difficult in this application due to the presence of a non-negligible mean. Using squared returns as a proxy for the unobservable conditional variance relies on expected returns being zero, an implausible assumption for investors holding risky assets. Engle and Colacito argue that if the measurement frequency is high enough, then expected returns are of a smaller order of magnitude than volatility, and so the mean can be ignored. In that case, the squared portfolio return, $\mathbf{w}_t'\mathbf{r}_t\mathbf{r}_t'\mathbf{w}_t$ or, better yet, the realised volatility of the portfolio, may be used as a volatility proxy and the estimated portfolio variance, $\mathbf{w}_t'\mathbf{H}_t\mathbf{w}_t$, can be evaluated using the univariate methods described in Sects. 2.1 and 2.4.

One important downside of minimum portfolio variance optimisation is that the mapping from $\mathbb{M}_{++}^K \rightarrow \mathbb{R}^K$ is many to one. There will generally be a continuum of forecasts that will produce the set of weights corresponding to the minimum variance portfolio. One direct method to address this deficiency is to test at least $K(K+1)/2$ portfolios using subsets of 2^K possible collections of assets.

4.2 Tracking error minimisation

Conditional covariance also plays a crucial role in estimating time-varying weights for tracking portfolios. Suppose a portfolio of $K-1$ stocks, with returns denoted $\widetilde{\mathbf{r}}_t$ is to be used track the return of another asset, r_t, by minimising the mean squared tracking error:

[16] If the returns are not computed as the difference between nominal returns and the risk-free rate, mean-variance analysis can still be used although with more complication. For details, see pp. 184–185 Campbell et al. (1997)

$$\min_{\mathbf{w}} \mathrm{E}[(r_t - \mathbf{w}'\widetilde{\mathbf{r}}_t)^2] \tag{62}$$

It is well known that the optimal linear projection satisfies

$$\mathbf{w} = E\left[\widetilde{\mathbf{r}}_t\widetilde{\mathbf{r}}_t'\right]^{-1} E[\widetilde{\mathbf{r}}_t r_t] \tag{63}$$

which are just the usual regression coefficients, see Hayashi (2000). Suppose the vector of returns is partitioned such that the first return, r_t, is the asset to be tracked so the covariance can be expressed

$$\mathbf{\Sigma}_t = \begin{bmatrix} \sigma^2_{1,t} & \mathbf{\Sigma}'_{12,t} \\ \mathbf{\Sigma}_{12,t} & \mathbf{\Sigma}_{22,t} \end{bmatrix}$$

where $V_{t-1}[r_t] = \sigma^2_{1,t}$ and $V_{t-1}[\widetilde{\mathbf{r}}_t] = \mathbf{\Sigma}_{22,t}$ The weights in the minimum tracking error portfolio can be computed by

$$\mathbf{w}_t = \mathbf{\Sigma}^{-1}_{22,t}\mathbf{\Sigma}_{12,t}. \tag{64}$$

Using the set of portfolio weights, the accuracy of the tracking portfolio weights can be tested with a MZ-GLS regression and the relative performance of two competing forecasts can be assessed using a Diebold-Mariano-West test. However, unlike the previously described minimum variance portfolio problems, it is not clear that utilising a forecast that satisfies H_0^* will lead to the smallest tracking error. However, the tracking error minimisation problem can be recast into the Engle-Colacito framework *if* the expected returns on all assets are assumed to be the same. In that case, the minimum variance portfolio problem becomes

$$\min_{\mathbf{w}_t} \ \mathbf{w}_t'\mathbf{\Sigma}_t\mathbf{w}_t \qquad \text{subject to } w_1 = 1$$

which can be re-written as the search for a global minimum variance portfolio by substituting in the constraint and expressing the returns on all assets as excesses above the return on the first asset.

4.3 Other methods of indirect evaluation

Volatility forecasts play a critical role in many other financial decisions, and accordingly their performance in these decisions is of much interest. Unfortunately, but perhaps not surprisingly, most financial decisions also depend on inputs beyond a volatility forecast. Influential assumptions about these other inputs will be required and these evaluations are "non-robust" using the definitions of this chapter. Nevertheless, we review the most common applications in this section.

One key application of volatility forecasts is in derivatives pricing, given the sensitivity of these securities to the volatility of the underlying asset price. These applications are generally univariate in nature; derivatives with multiple underlying assets (thus requiring a covariance forecast) are much less studied in the literature (see Bates (2003), Garcia et al. (2008) for surveys). In any applications to derivatives, the volatility forecast must be combined with a pricing model – the most commonly-used such model is the Black-Scholes model for European options. If the pricing model is mis-specified then the ranking of two volatility forecasts by pricing errors will not necessarily lead to the true conditional variance being ranked above imperfect forecasts. However if the volatility forecast *is* to be used in a particular pricing model, then the interest of the forecast user is not necessarily in finding the true conditional variance, rather it is in finding the forecast that produces the smallest pricing errors, and so this "distortion" is not a cause for concern. The evaluation and comparison of volatility forecasts via derivative pricing problems has been previously considered (see Noh et al. (1994), Gibson and Boyer (1997), Christoffersen and Jacobs (2004b), Christoffersen and Jacobs (2004a), Gonźalez-Rivera et al. (2004)).

Another important financial application involving volatility forecasts is portfolio decisions. We reviewed two special cases in the previous section where only the covariance was required for a portfolio decision. For most utility functions and general returns distributions, other inputs are needed to choose portfolio weights, and the specification of these inputs can affect the ranking of volatility forecasts by the out-of-sample utility of portfolio returns. Applications of volatility forecasts in portfolio decisions are have been widely explored (see West et al. (1993), Fleming et al. (2003), Marquering and Verbeek (2004), Gonźalez-Rivera et al. (2004), amongst others).

In the past decade or so, measures of risk beyond standard deviation have gained increasing attention. Most prominent amongst these is Value-at-Risk (VaR), which can be defined as the α-quantile of the conditional distribution of the return on a given asset or portfolio[17]. In most applications, α is set to 0.05 or 0.01. See Christoffersen (2008) in this Handbook for more on VaR. It is possible to produce conditional VaR forecasts by directly modelling the conditional quantile of interest, see Engle and Manganelli (2004). However, the majority of conditional VaR forecasts are produced by first specifying a model for the conditional variance, and then specifying a model for the quantile of the standardised residual. This link between conditional VaR forecasts and conditional variance forecasts has lead some authors to suggest testing variance forecasts by testing the performance of VaR forecasts based on the variance forecast(s), although mis-specification of the distribution for the standardised residuals can lead to the rejection of a forecast satisfying H_0^*. VaR-based evaluation and comparison of volatility forecast has been ex-

[17] Another prominent measure of risk is "expected shortfall", which is defined as the conditional mean of the return on an asset given that the return is less than the Value-at-Risk, ie: $ES_t \equiv E\left[r_t | \mathcal{F}_{t-1} \cap r_t \leq VaR_t\right]$.

amined in Lopez (2001), Gonźalez-Rivera et al. (2004), Ferreira and Lopez (2005), Kuester et al. (2006), amongst others.

5 Conclusion

This chapter provided an overview of the methods available for evaluating and comparing forecasts of the conditional variance of an asset return or the conditional covariance of a set of asset returns. We paid particular attention to the problems that arise due to the fact that volatility is unobservable. We emphasised the importance of using tests (a) that are robust to the "noise" in the volatility proxy, if a proxy used, and (b) that require only minimal assumptions on the distribution of the returns. Many widely-employed volatility forecast evaluation tests, such as those using VaR or option pricing, fail one or both of these criteria.

In addition to presenting the theory for methods that satisfy these two criteria, we also presented the results of some small Monte Carlo studies to provide guidance for empirical work. We suggested a modification of the widely-used Mincer-Zarnowitz regression for testing volatility forecast optimality, which exploits the additional structure that holds under the null hypothesis. Our suggested "MZ-GLS" test has good size and much better power in finite samples than other MZ tests. Our simulations also clearly demonstrated the value of higher-precision volatility proxies, such as realised variance (cf. Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2004)). Even simple estimators based on 30-minute returns provide large gains in power and improvements in finite-sample size.

In Monte Carlo studies using realistic stochastic volatility processes, we studied the choice of loss function in Diebold-Mariano-West (DMW) tests, see Diebold and Mariano (1995) and West (1996). We showed that the use of loss functions that are "non-robust", in the sense of Patton (2006), can yield perverse rankings of forecasts, even when very accurate volatility proxies are employed. Amongst the class of robust and homogeneous loss functions in Patton and the multivariate generalisation of these loss functions provided in this chapter, our small simulations suggested that the "QLIKE" loss functions yield the greatest power in DMW tests.

References

Andersen, T.G. and Bollerslev, T. (1998): Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**, 885–905.

Andersen, T.G., Benzoni, L. and Lund, J. (2002): An empirical investigation of continuous-time equity return models. *Journal of Finance* **57**, 1239–1284.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003): Modeling and Forecasting Realized Volatility. *Econometrica* **71**, 3–29.

Andersen, T.G., Bollerslev, T. and Meddahi, N. (2005): Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica* **73**, 279–296.

Andersen, T.G., Bollerslev, T., Christoffersen, P.F. and Diebold, F.X. (2006a): Volatility and correlation forecasting. *In: Elliott G., Granger C., Timmermann A. (Eds.): Handbook of Economic Forecasting*. North Holland, Amsterdam.

Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2006b): Parametric and nonparametric volatility measurement. *In: Hansen L.P., Aït-Sahalia Y. (Eds.): Handbook of Financial Econometrics*, forthcoming. North-Holland, Amsterdam.

Arellano, M. (2003): *Panel Data Econometrics*. Oxford University Press, Oxford.

Barndorff-Nielsen, O.E. and Shephard, N. (2004): Econometric analysis of realised covariation: high frequency based covariance, regression and correlation in financial economics. *Econometrica* **73**, 885–925.

Bates, D.S. (2003): Empirical option pricing: a retrospection. *Journal of Econometrics* **116**, 387–404.

Bierens, H.J. (1990): A consistent conditional moment test of functional form. *Econometrica* **58**, 1443–1458.

Bierens, H.J. and Ploberger, W. (1997): Asymptotic theory of integrated conditional moment tests. *Econometrica* **65**, 1129–1152.

Bollerslev, T. and Wright, J.H. (2001): High-frequency data, frequency domain inference, and volatility forecasting. *The Review of Economics and Statistics* **83**, 596–602.

Bollerslev, T., Engle, R.F. and Wooldridge, J.M. (1988): A capital asset pricing model with time-varying covariances. *Journal of Political Economy* **96**, 116–131.

Campbell, J.Y., Lo, A.W. and MacKinlay, A.C. (1997): *The Econometrics of Financial Markets*. Princeton University Press, Princeton.

Chernov, M. (2007): On the role of risk premia in volatility forecasting. *Journal of Business and Economic Statistics* forthcoming.

Chernov, M., Gallant, A.R., Ghysels, E. and Tauchen, G. (2003): Alternative models for stock price dynamics. *Journal of Econometrics* **116**, 225–257.

Chib, T. (2008): Multivariate stochastic volatility. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 365–400. Springer, New York.

Christensen, K. and Podolskij, M. (2006): Realized range-based estimation of integrated variance. *Journal of Econometrics* forthcoming.

Christodoulakis, G.A. and Satchell, S.E. (2004): Forecast evaluation in the presence of unobserved volatility. *Econometric Reviews* **23**, 175–198.

Christoffersen, P. (2008): Estimation of value-at-risk. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 752–766. Springer, New York.

Christoffersen, P.F. and Jacobs, K. (2004a): The importance of the loss function in option valuation. *Journal of Financial Economics* **72**, 291–318.

Christoffersen, P.F. and Jacobs, K. (2004b): Which Garch model for option valuation? *Management Science* **50**, 1204–1221.

Diebold, F.X. and Mariano, R.S. (1995): Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**, 253–263.

Engle, R.F. and Colacito, R. (2006): Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics* **24**, 238–253.

Engle, R.F. and Manganelli, S. (2004): Caviar: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* **22**, 367–381.

Ferreira, M.A. and Lopez, J.A. (2005): Evaluating interest rate covariance models within a value-at-risk framework. *Journal of Financial Econometrics* **3**, 126–168.

Fleming, J., Kirby, C. and Ostdiek, B. (2003): The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics* **67**, 473–509.

Garcia, R., Ghysels, E. and Renault, E. (2008): The econometrics of option pricing. *In: Aït-Sahalia Y., Hansen L.P. (Eds.): Handbook of Financial Econometrics*, forthcoming. Elsevier-North Holland, Amsterdam.

Giacomini, R. and White, H. (2006): Tests of conditional predictive ability. *Econometrica* **74**, 1545–1578.

Gibson, M.S. and Boyer, B.H. (1997): *Evaluating forecasts of correlation using option pricing, board of Governors of the Federal Reserve System* (U.S.).

Gonçalves, S. and Meddahi, N. (2005): Bootstrapping realized volatility, *Département de Sciences Économiques, CIREQ and CIRANO Université de Montréal.*

Gonźalez-Rivera, G., Lee, T.H. and Mishra, S. (2004): Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting* **20**, 629–645.

Griffin, J.E. and Oomen, R.C. (2006): *Covariance measurement in the presence of non-synchronous trading and market microstructure noise. Mimeo.*

Hansen, P.R. (2005): A test for superior predictive ability. *Journal of Business and Economic Statistics* **23**, 365–380.

Hansen, P.R. and Lunde, A. (2005): A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* **20**, 873–889.

Hansen, P.R. and Lunde, A. (2006a): Consistent ranking of volatility models. *Journal of Econometrics* **127**, 97–121.

Hansen, P.R. and Lunde, A. (2006b): Realized variance and market microstructure noise. *Journal of Business and Economic Statistics* **24**, 127–218.

Hansen, P.R., Lunde, A. and Nason, J.M. (2005): Model confidence sets for forecasting models, *Federal Reserve Bank of Atlanta, Working Paper* **7**.

Harvey, A., Ruiz, E. and Shephard, N. (1994): Multivariate stochastic variance models. *Review of Economic Studies* **61**, 247–264.

Hayashi, F. (2000): *Econometrics.* Princeton University Press.

de Jong, R.M. (1996): The Bierens test under data dependence. *Journal of Econometrics* **72**, 1–32.

Jorion, P. (1995): Predicting volatility in the foreign exchange market. *Journal of Finance* **50**, 507–528.

Koopman, S.J. (2008): Parameter estimation and practical aspects of modelling stochastic volatility. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 312–344. Springer, New York.

Kuester, K., Mittnik, S. and Paolella, M.S. (2006): Value-at-Risk Prediction: A Comparison of Alternative Strategies. *Journal of Financial Econometrics* **4**, 53–89.

Lopez, J.A. (2001): Evaluating the predictive accuracy of volatility models. *Journal of Forecasting* **20**, 87–109.

Magnus, J.R. and Neudecker, H. (2002): *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, Chichester.

Marquering, W. and Verbeek, M. (2004): The economic value of predicting stock index returns and volatility. *Journal of Financial and Quantitative Analysis* **39**, 407–429.

Meddahi, N. (2001): A theoretical comparison between integrated and realized volatilities. *Manuscript, Département de Sciences Économiques, CIREQ and CIRANO Université de Montréal.*

Mincer, J. and Zarnowitz, V. (1969): The evaluation of economic forecasts. *In: Mincer J. (Ed.): Economic Forecasts and Expectations.* Columbia University Press.

Newey, W.K. and West, K.D. (1987): A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.

Noh, J., Engle, R.F. and Kane, A. (1994): Forecasting volatility and option prices of the S&P 500 index. *Journal of Derivatives* **2**, 17–30.

Parkinson, M. (1980): The extreme value method for estimating the variance of the rate of return. *The Journal of Business* **53**, 61–65.

Patton, A.J. (2006): Volatility forecast comparison using imperfect volatility proxies, *Quantitative Finance Research Centre*, University of Technology Sydney, Research Paper **175**.

Sheppard, K. (2006): *Realized covariance and scrambling* Univeristy of Oxford.

Silvennoinen, A., Teräsvirta, T. (2008): Multivariate Garch models. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 201–229. Springer, New York.

Teräsvirta, T. (2008): An introduction to univariate GARCH. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 17–42. Springer, New York.

Theil, H. (1958): *Economic Forecasts and Policy*. North-Holland, Rotterdam.

Tse, Y.K. (2000): A test for constant correlations in a multivariate GARCH model. *Journal of Econometrics* **98**, 107–127.

West, K.D. (1996): Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–1084.

West, K.D. (2006): Forecast evaluation. *In: Elliott G., Granger C., Timmermann A. (Eds.): Handbook of Economic Forecasting*. North Holland, Amsterdam.

West, K.D., Edison, H.J. and Cho, D. (1993): A utility-based comparison of some models of exchange rate volatility. *Journal of International Economics* **35**, 23–45.

White, H. (1980): A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

White, H. (1996): Estimation, Inference and Specification Analysis. *Econometric Society Monographs*. Cambridge University Press.

White, H. (2000): A reality check for data snooping. *Econometrica* **68**, 1097–1126.

Zivot, E. (2008): Practical aspects of Garch-modelling. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 112–155. Springer, New York.

Structural Breaks in Financial Time Series

Elena Andreou and Eric Ghysels*

Abstract This paper reviews the literature on structural breaks in financial time series. The second section discusses the implications of structural breaks in financial time series for statistical inference purposes. In the third section we discuss change-point tests in financial time series, including historical and sequential tests as well as single and multiple break tests. The fourth section focuses on structural break tests of financial asset returns and volatility using the parametric versus nonparametric classification as well as tests in the long memory and the distribution of financial time series. In concluding we provide some areas of future research in the subject.

1 Introduction

There are statistical inference as well as investment allocation implications of ignoring structural changes in financial processes. On statistical inference grounds, it is shown that ignoring structural breaks in financial time series can yield spurious persistence in the conditional volatility. For instance, neglected structural changes can yield Integrated GARCH or long memory effects in financial time series (e.g., Diebold (1986), Lamoureux and Lastrapes (1990), Mikosch and Stărică (2004), Hillebrand (2005)) and can have implications about the existence of higher order unconditional moments such as the kur-

Elena Andreou
University of Cyprus, Department of Economics, P.O. Box 537, CY 1678 Nicosia, Cyprus, e-mail: elena.andreou@ucy.ac.cy

Eric Ghysels
Department of Economics, Gardner Hall, CB 3305, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3305, e-mail: eghysels@unc.edu

* We would like to thank Richard A. Davis and Stacey Hancock for carefully reading the chapter and for their valuable suggestions.

DOI: 10.1007/978-3-540-71297-8_37,

tosis or the tail index in financial time series (e.g., Andreou and Ghysels (2005), Mikosch and Stărică (2004)) as well as forecasting (e.g., Pesaran and Timmerman (2004), Pesaran et al. (2006)).

From an economic perspective, there is empirical evidence showing that there are structural breaks in financial markets which affect fundamental financial indicators. Examples of these indicators are financial returns and volatility (e.g., Lamoureux and Lastrapes (1990), Andreou and Ghysels (2006a), Horváth et al. (2006)), the shape of the option implied volatility smile (Bates (2000)), asset allocation (e.g., Pettenuzzo and Timmerman (2005)), the equity premium (Pastor and Stambaugh (2001), Chang-Jin et al. (2005)), the tail of the distribution and risk management measures such as Value at Risk (VaR) and Expected Shortfall (ES) (Andreou and Ghysels (2005)) as well as credit risk models and default measures (Andreou and Ghysels (2007)). Finally, empirical evidence shows that various economic events can lead to structural changes detected in a large number of financial series, such as the financial liberalization of emerging markets and integration of world equity markets (see, for instance, Garcia and Ghysels (1998), Bekaert, Harvey and Lumsdaine (2002) inter alia), changes in exchange rate regimes such as the collapse of exchange rate systems (e.g., the Exchange Rate Mechanism) and the introduction of a single currency in Europe.

The topics addressed in this review are the following: Section 2 discusses some statistical inference implications of ignoring change-points in financial processes. Section 3 starts with a brief review of the properties of financial time series, focusing on their temporal dependence and stationarity assumptions. The assumptions on strong mixing are satisfied by a large class of financial series which are going to be the basis for discussing the asymptotic properties of structural breaks tests. We then discuss change-point tests in financial time series such as historical and sequential tests as well as single and multiple break tests. The fourth section discusses structural change tests in returns, volatility, long memory and distribution of financial time series. The paper concludes with some areas of future research.

2 Consequences of Structural Breaks in Financial Time Series

This section discusses some of the statistical inference implications of structural breaks on the persistence of financial time series, such as long range dependence and integrated GARCH effects. Assume that financial returns, r_t, follow a discrete time GARCH(1,1) process, which captures many of the stylized facts of financial time series and is used in many studies as the benchmark model, given by:

$$r_t = \sigma_t u_t, \quad \sigma_t = \alpha_0 + a_1 \sigma_{t-1} + d_1 r_{t-1}^2, \quad t = 1, 2, ..., \quad u_t \sim i.i.d(0, 1). \quad (1)$$

Although (1) can be considered as a very simple mechanism that may not capture all the stylized facts of financial returns, it nevertheless suffices for discussing the implications of structural breaks in financial time series.

First, we focus on structural breaks in the unconditional variance of financial processes and explain how these can yield spurious long-memory effects and an integrated GARCH (IGARCH), $a_1 + d_1 = 1$ in (1). In the general context, a second order stationary sequence Y_t is said to exhibit long memory if the condition $\sum_h |\rho_Y(h)| = \infty$ holds, where $\rho_Y(h) = corr(Y_0, Y_h)$, $h \in \mathcal{Z}$, denotes the ACF of the Y_t sequence. Alternatively, the long range dependence via the power law decay of the autocorrelation function is given by: $\rho_Y(h) \sim c_\rho h^{2d-1}$ for a constant $c_\rho > 0$, for large h and some $d \in (0, 0.5)$. Mikosch and Stărică (2004) show how statistical tools like the sample ACF and periodogram of a process (1) can yield long-range effects when there are unaccounted nonstationarities such as shifts in the mean or variance.

When there are multiple change-points, the sample $Y_1, ..., Y_T$ consists of different subsamples from distinct stationary models. To be precise, let p_j, $j = 0, ..., r$ be positive numbers such that $p_1 + ... + p_r = 1$ and $p_0 = 0$. Define $q_j = p_0 + ... + p_j$, $j = 0, ..., r$. The sample $Y_1, ..., Y_T$ is written as

$$Y_1^{(1)}, ..., Y_{[Tq_1]}^{(1)}, ..., Y_{[Tq_{r-1}]+1}^{(r)}, ..., Y_T^{(r)} \tag{2}$$

where the r subsamples come from distinct stationary ergodic models with finite second moment. Given the nonstationary sample (2), the sample autocovariances of the sequence Y_t are given by $\widetilde{\gamma}_{T,Y}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(Y_t - \overline{Y}_T)(Y_{t+h} - \overline{Y}_T)$, $h \in \mathcal{T}$. By the ergodic theorem, it follows for fixed $h \geq 0$ as $T \to \infty$ that

$$\begin{aligned}
&\widetilde{\gamma}_{T,Y}(h) \\
&= \sum_{j=1}^{r} \frac{p_j}{Tp_j} \sum_{t=[Tq_{j-1}]+1}^{[Tq_j]} Y_t^{(j)} Y_{t+h}^{(j)} - \left(\sum_{j=1}^{r} \frac{p_j}{Tp_j} \sum_{t=[Tq_{j-1}]+1}^{[Tq_j]} Y_t^{(j)} \right)^2 + o(1) \\
&\to \sum_{j=1}^{r} p_j E\left(Y_0^{(j)} Y_h^{(j)} \right) - \left(\sum_{j=1}^{r} p_j E Y^{(j)} \right)^2 \\
&= \sum_{j=1}^{r} p_j \gamma_{Y^{(j)}}(h) - \sum_{1 \leq i < j \leq r} p_i p_j \left(EY^{(j)} - EY^{(i)} \right)^2 \quad a.s.
\end{aligned} \tag{3}$$

Let r_t follow a GARCH type model such as that given in (1). If $Y_t = |r_t|$ or r_t^2 in (3) the expectations of subsequences $Y_t^{(j)}$ differ, and since the sample autocovariances $\widetilde{\gamma}_{Y^{(j)}}(h)$ within each stationary segment decay to zero exponentially as $h \to \infty$ (due to the short memory assumption), the sample ACF $\widetilde{\gamma}_{T,Y}(h)$ for sufficiently large h is close to a strictly positive constant given by the last term in (3). The shape of a sample ACF ($\widetilde{\gamma}_{T,Y}(h)$) decays ex-

ponentially for small lags and approaches a positive constant for larger lags. Hence the samples of $|r_1|, ..., |r_n|$ and $r_1^2, ..., r_n^2$ have sample ACFs that decay quickly for the first lags and then they approach positive constants given by:

$$\sum_{1 \le i < j \le r} p_i p_j \left(E|r^{(j)}| - E|r^{(i)}| \right)^2$$

and

$$\sum_{1 \le i < j \le r} p_i p_j \left(E\left(r^{(j)}\right)^2 - E\left(r^{(i)}\right)^2 \right)^2, \tag{4}$$

respectively, which would explain the long memory observed in financial returns (Mikosch and Stărică (2004)). Moreover, the stronger the nonstationarity which implies a bigger difference in (3), the more pronounced the long-memory effect in the ACF. Mikosch and Stărică (2004) also show that the Whittle estimate of the ARMA representation of the GARCH model will be close to unity when there are unaccounted change-points.

The spurious IGARCH effects due to unaccounted structural breaks or parameter regime switches in financial processes have been documented early in the empirical literature (e.g., Diebold (1986), Lamoureux and Lastrapes (1990)). More recently, Hillebrand (2005) and Mikosch and Stărică, (2004) provide a theoretical explanation for this effect. In particular, Hillebrand (2005) shows that unaccounted structural breaks in the unconditional variance yield a spurious IGARCH which is a consequence of the geometry of the estimation problem. This result is independent of the estimation method and the statistical properties of parameter changes and generalizes to higher order GARCH models. Consider, for example, the single change-point in the conditional variance parameters of a GARCH. In each of the two segments, the realizations of the conditional volatility process are centered approximately around the unconditional, stationary mean corresponding to the parameters of that segment. If the GARCH model is estimated globally without accounting for segmentation, the resulting hyperplane (parameterized by $\widehat{\alpha}_0, \widehat{a}_1, \widehat{d}_1$) must go through both segment means of σ_t. As the mean of σ_t and the mean of σ_{t-1} is the same for sufficiently long segments, a line connecting two different means in the $(\sigma_t, \sigma_{t-1})-$subspace is close to the identity. Therefore, the estimator of a_1 will pick up the slope of the identity and be close to one. The remaining autoregressive parameter d_1 will be chosen residually such that $\widehat{a}_1 + \widehat{d}_1 \approx 1$ causes the spurious IGARCH effect. The sum will always stay slightly below one to keep the estimated volatility process from exploding. This is proved in Hillebrand (2005).

The bias problem in the persistence of the volatility of financial time series is more serious than in linear AR processes when breaks are ignored. This is because the source of stochasticity in the GARCH equation originates from r_{t-1}^2 alone, and there is no contemporaneous error that is orthogonal to the regressors r_{t-1}^2 and σ_{t-1}. Hence, in GARCH models one does not find the

interplay of the distance in the conditional local means which is given by $\alpha_0/(1 - a_1 - d_1)$, with the variance of the orthogonal error process as is the case with the AR model. Consequently, GARCH models are more sensitive to change-points than linear AR models.

The spurious IGARCH effects imply infinite unconditional variance. This has financial theory implications given that many asset pricing models are based on the mean-variance theorem. In addition, it also has statistical inference implications such as, for instance, for forecasting since it implies that shocks have a permanent effect on volatility such that current information remains relevant when forecasting the conditional variance at all horizons. A related strand of literature on forecasting and structural breaks that is relevant for long horizon financial returns captured by linear models (with no dynamic volatility effects) can be found, for instance, in Pesaran and Timmerman (2004). They show analytically that it is costly to ignore breaks when forecasting the sign or direction of a time-series subject to a large break, and a forecasting approach that conditions on the most recent break is likely to perform better over unconditional approaches that use expanding or rolling estimation windows. Further investigation of these results in the context of the rolling volatility estimators when there is stochastic volatility as in Foster and Nelson (1996) and when there are breaks is still unexplored. Last but not least, structural breaks can yield misspecification in the asymmetry and the tails of the conditional distribution of returns (Andreou and Ghysels (2005)).

3 Methods for Detecting Structural Breaks

In this section we review statistical methods for detecting change points in financial processes. One classification of change-point tests refers to the distinction between a posteriori, retrospective or historical tests versus sequential, a priori or on-line tests. These two methods are classified according to the sample acquisition approach. For the a posteriori change-point tests, the process of data acquisition is completed at the moment when the homogeneity hypothesis is checked while for sequential structural break tests, this hypothesis is tested on-line with observations, i.e., simultaneously with the process of data acquisition. Hence, the sequential approach is particularly useful when a decision has to be made on–line, as new data become available. Although sequential tests were originally introduced in order to construct more efficient inspection procedures for industrial processes, they can also be useful for financial processes and especially for financial decision making such as risk management, asset allocation and portfolio selection.

In the first subsection, we discuss the general assumptions of financial returns underlying the statistical procedures. The next subsection discusses

historical and sequential methods for detecting breaks. We then turn to multiple change-point detection methods.

3.1 Assumptions

We denote a generic process by r_t that represents the financial asset returns. Under the null hypothesis of no structural change, we assume that r_t (i) is a weakly stationary process with uniformly bounded $(2+\delta)th$ moments for some $0 < \delta \leq 2$ and (ii) is a strong mixing process. Then, letting $Y_T = r_1 + ... + r_T$, the limit $\sigma_Y^2 = \lim_{T\to\infty} \frac{1}{T} EY_T^2$ exists, and if $\sigma_Y > 0$, then there exists a Wiener process $\{W(t), 0 \leq t < \infty\}$ such that $Y_T - \sigma W(T) = O(T^{1/2-\varepsilon})$ a.s. where $\varepsilon = \delta/600$ (see for instance, Phillip and Stout (1975), Theorem 8.1, p. 96). This result is general enough to cover many applications.

Under the above mixing and stationarity conditions, the process r_t satisfies the strong invariance principle:

$$\sum_{1\leq t\leq T} (r_t - E(r_t)) - \sigma_Y W(T) = o(T^{\gamma}), \tag{5}$$

a.s. with some $0 < \gamma < 0.5$ and $W(.)$ a Wiener process. Consequently, under the null of no structural change, $Y_t = |r_t|^{\upsilon}, \upsilon = 1, 2$ satisfies the Functional Central Limit Theorem (FCLT)

$$Z_T := T^{-1/2} \sum_{1\leq t\leq T} (r_t - E(r_t)) \to \sigma_Y W(T), \tag{6}$$

for a large sample size, T.

The above conditions are satisfied by the continuous and discrete time models for financial returns. For example, for the Heston model and other stochastic volatility (SV) models Genon-Catalot et al. (2000), Proposition 3.2, p. 1067, show that the volatility model is $\beta-$mixing (which implies $\alpha-$mixing). The key insight of Genon-Catalot et al. (2000) is that continuous time SV models can be treated as hidden Markov processes when observed discretely which thereby inherit the ergodicity and mixing properties of the hidden chain. Carrasco and Chen (2002) extend this result to generalized hidden Markov chains and show $\beta-$mixing for the SV-AR(1) model (Andersen (1994)). Other SV specifications found in Chernov et al. (2003) also satisfy the $\beta-$mixing condition. In addition, Carrasco and Chen (2002) and Davis and Mikosch (1998) show that discrete time models for financial returns, such as, for instance, the family of GARCH models also satisfy $\beta-$mixing.

3.2 Historical and sequential partial–sums change–point statistics

For conciseness and for comparison with sequential statistics, we focus on a CUSUM test for breaks, which is one of the most popular change-point tests. Although this test was originally developed for independent processes for detecting a break in the mean (e.g., Page (1955)) or the variance (e.g., Inclan and Tiao (1994)) it has recently been extended to β-mixing processes (e.g., Kokoszka and Leipus (2000)) for detecting change-points in an ARCH type process.

Let the asset returns process, r_t, be a β-mixing process with finite fourth moment. A large class of ARCH and SV models are β-mixing that satisfy the assumptions described in the previous section. Define the process of interest $Y_t = |r_t|^\delta$ for $\delta = 1, 2$, which represents an observed measure of the variability of returns. For $\delta = 2$ the squared returns is the parent process parametrically modeled in ARCH- or SV-type models. Alternatively, when $\delta = 1$, absolute returns, is considered as another measure of risk, in, say, the Power-ARCH models. Given that the measurable functions of mixing processes are also mixing and of the same size (see White (1984), Theorem 3.49) then $Y_t = G(r_t, \ldots, r_{t-\tau})$, for finite τ, defined by $Y_t = |r_t|^\delta$ for $\delta = 1, 2$, is also β-mixing. The tests discussed in this section will examine whether there is evidence of structural breaks in the dynamics of stock returns volatility, which is one of the moments of interest in financial processes. Note that these tests would not necessarily require the specification of the functional form of volatility. Andreou and Ghysels (2002) extend this analysis to sampling returns intra-daily, denoted $r_{(i),t}$ for some intra-day frequency $i = 1, \ldots, m$, and form data-driven estimates of daily volatility by taking sums of squared intra-day returns. This is an example of $Y_t = G(r_{(1),t}, \ldots, r_{(m),t})$. The high frequency process is β-mixing, and so is the daily sampled sum of intra-day squared returns, or various other empirical measures of Quadratic Variation (QV). For example, $Y_t := (QVi)_t$ are locally smoothed filters of the quadratic variation using i days of high-frequency data. The case of $QV1$ corresponds to the filters studied by Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2001).

In order to test for breaks in an ARCH(∞), Kokoszka and Leipus (2000) consider the following process:

$$U_T(k) = \left(\frac{k(T-k)}{T^2}\right)^{1/2} \left(\frac{1}{k}\sum_{j=1}^{k} Y_j - \frac{1}{T-k}\sum_{j=k+1}^{T} Y_j\right), \tag{7}$$

where $0 < k < T$, $Y_t = r_t^2$. The returns process $\{r_t\}$ follows an ARCH(∞) process, $r_t = u_t\sqrt{\sigma_t}$, $\sigma_t = a + \sum_{j=1}^{\infty} b_j r_{t-j}^2$, $a \geq 0, b_j \geq 0, j = 1, 2$, with finite fourth moment and errors u_t that can be non-Gaussian. An alternative way of expressing (7) is:

$$U_T(k) = \left(\frac{1}{\sqrt{T}} \sum_{j=1}^{k} Y_j - \frac{k}{T\sqrt{T}} \sum_{j=1}^{T} Y_j \right). \tag{8}$$

The CUSUM type estimator $\widehat{k}$ of a change point k^* is defined as:

$$\widehat{k} = \min \left\{ k : |U_T(k)| = \max_{1 \le j \le T} |U_T(j)| \right\}. \tag{9}$$

The estimate $\widehat{k}$ is the point at which there is maximal sample evidence for a break in the squared returns process. In the presence of a single break, it is proved that $\widehat{k}$ is a consistent estimator of the unknown change-point k^* with $P\{|k^* - \widehat{k}| > \varepsilon\} \le C/(\delta \varepsilon^2 \sqrt{T})$, where C is some positive constant, δ depends on the ARCH parameters and $|k^* - \widehat{k}| = O_p(T^{-1})$. Using the FCLT for $\beta-$mixing processes, it is possible to show that under the null hypothesis of no break:

$$U_T(k) \rightarrow_{D[0,1]} \sigma B(k), \tag{10}$$

where $B(k)$ is a Brownian bridge and $\sigma^2 = \sum_{j=-\infty}^{\infty} Cov(Y_j, Y_0)$. Consequently, using an estimator $\widehat{\sigma}$, one can establish that under the null:

$$\sup\{|U_T(k)|\}/\widehat{\sigma} \rightarrow_{D[0,1]} \sup\{B(k) : k \epsilon [0,1]\}, \tag{11}$$

which requires a Heteroskedasticity and Autocorrelation Consistent (HAC) estimator applied to the Y_j process.

We can relate the Kokoszka and Leipus (2000) statistic (7) to that of Inclan and Tiao (IT) (1994). The IT CUSUM test statistic for detecting a break in the variance of an independent process is:

$$IT = \sqrt{T/2} \max_k |D_k|, \tag{12}$$

where

$$D_k = \left[\left(\sum_{j=1}^{k} Y_j / \sum_{j=1}^{T} Y_j \right) - k/T \right].$$

This is related to (7) as follows:

$$U_T(k) = \left(\frac{1}{k(T-k)} \right)^{1/2} \left(\frac{1}{T} \sum_{j=1}^{k} Y_j \right) D_k. \tag{13}$$

The above tests can also be used to assess change-points in a general location-scale model given by

$$r_t = \mu(Z_t) + \sigma(Z_t)\varepsilon_t, \tag{14}$$

where $\{(r_t, Z_t), t = 1, 2, ...\}$ is a sequence of random variables, $\{\varepsilon_t\}$ is a sequence of stationary errors with $E(\varepsilon_t/Z_t) = 0$ and $var(\varepsilon_t/Z_t) = 1$, and $\mu(Z_t)$

and $\sigma(Z_t)$ are the conditional mean and skedastic functions, respectively. ARCH-type models correspond to $\mu(Z_t) = 0$ and appropriate functional forms for $\sigma(Z_t)$. Chen et al. (2005) establish the asymptotic properties of estimators for structural breaks in volatility when the regression and skedastic functions are unknown but estimated nonparametrically via local polynomial (linear) smoothers by proposing new methods to select the bandwidths. Their statistic is a CUSUM-type test given in (7) with the same Brownian Bridge asymptotic distributions, but their monitoring process is now the nonparametric residual, $Y_t = (r_t - \widehat{\mu}(Z_t))/\widehat{\sigma}(Z_t)$. Their change-point estimator is consistent and also converges with a rate of $O(T^{-1})$.

Other types of partial-sums tests can also be used to detect breaks in GARCH models as well as the Lagrange Multiplier tests found in Chu (1995) and in Lundbergh and Teräsvirta (1998).

We now turn to sequential change-point tests. Sequential test statistics compare the process over a historical sample $1, ..., m$ $(m < T)$ with the process over the monitoring sample $m+1, m+2, ...$ The historical sample represents where the process is in-control or noncontaminated by a break. Sequential change-point tests for financial time series have some important advantages since sampling does not involve any significant cost and has implications for the power of the tests (Andreou and Ghysels (2006a)).

The following sequential partial-sums type test statistics, S_T, are considered for monitoring the process Y_t, which is a function of the financial returns process. The Fluctuation (FL) detector is given by:

$$S_T^{FL} = (T-m)\widehat{\sigma_0}^{-1}(\overline{Y}_{T-m} - \overline{Y}_m), \tag{15}$$

$$\overline{Y}_{T-m} = \frac{1}{T-m}\sum_{j=m+1}^{T} Y_j,$$

measures the updated mean estimate, $\overline{Y}_{T-m}$, from the historical mean estimate, $\overline{Y}_m$ and $\widehat{\sigma_0}$ is the variance estimator from the historical sample. The Partial Sums (PS) statistic:

$$S_T^{PS} = \sum_{i=m+1}^{m+k} (Y_i - \overline{Y}_m), k \geq 1 \tag{16}$$

is similar to the Cumulative Sums (CUSUM) test of Brown et al. (1975) in that it monitors the least squares residuals $Y_i - \overline{Y}_m$. The Page (PG) CUSUM statistic is:

$$S_T^{PG} = \sum_{i=1}^{T} Y_i - \min_{1 \leq i < T} \sum_{i=1}^{T} Y_i \tag{17}$$

which can also be considered as a Partial Sums type test since for an independent Y_i, it is equivalent to $\sum_{i=1}^{T} Y_i - \sum_{i=1}^{T-r} Y_i$ for any r, $1 \leq r \leq n$ (see Page (1954)). The asymptotic distribution of the above sequential statistics

can be derived using the framework of Kuan and Hornik (1995) and Leisch et al. (2000). A detailed discussion of the asymptotic results of the above sequential change-point statistics as well as the bounderies associated with these statistics can be found in Andreou and Ghysels (2006b).

Other sequential change-point tests based on the quasi-likelihood function of the volatility of financial time series models can be found in Berkes et al. (2004).

3.3 Multiple breaks tests

This section discusses multiple change-point detection methods in financial time series. The tests considered here assume unknown breaks in the variance of a temporally dependent process. We divide the multiple breaks methods into two categories: those based on the model selection approach and those based on the binary, sequential segmentation of the sample.

The challenge in multiple change-point testing is to jointly estimate the location of the breaks and the corresponding length of segments or regimes between breaks, while also providing estimates of the model parameters and possibly orders of the time series model in each segment. To formalize the problem, consider the process $\{Y_t\}$ characterized by a parameter $\theta \in \Theta$ that remains constant between subsequent changes. Consider the set of change-points $\tau = \{\tau_1, \tau_2, ..., \tau_{K-1}\}$ where K defines an integer and $0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < T$, where $\tau_0 = 0$ and $\tau_K = T$. For any $1 \leq k \leq K$ use the contrast function $U(Y_{\tau_{k-1}+1}, ..., Y_{\tau_k}; \theta)$ useful for the estimation of the unknown true value of the parameter in the segment or regime k. The minimum contrast estimate $\widehat{\theta}(Y_{\tau_{k-1}+1}, ..., Y_{\tau_k})$, computed on segment k of τ, is defined as a solution to the following minimization problem:

$$U(Y_{\tau_{k-1}+1}, ..., Y_{\tau_{k-1}}; \widehat{\theta}(Y_{\tau_{k-1}+1}, ..., Y_{\tau_{k-1}})) \leq U(Y_{\tau_{k-1}+1}, ..., Y_{\tau_{k-1}}; \theta), \forall \theta \in \Theta.$$

For any $1 \leq k \leq K$, let G be defined as:

$$G(Y_{\tau_{k-1}+1}, ..., Y_{\tau_k}) = U(Y_{\tau_{k-1}+1}, ..., Y_{\tau_k}; \widehat{\theta}(Y_{\tau_{k-1}+1}, ..., Y_{\tau_k})).$$

Then, define the contrast function $J(\tau, \mathbf{Y})$ as:

$$J(\tau, \mathbf{Y}) = \frac{1}{T} \sum_{k=1}^{K} G(Y_{\tau_{k-1}+1}, ..., Y_{\tau_k}).$$

In the case of detecting changes in the variance of a sequence of random variables, the following contrast function, based on the Gaussian log-likelihood function, can be used:

$$J(\tau, \mathbf{Y}) = \frac{1}{T} \sum_{k=1}^{K} T_k \log(\widehat{\sigma}_k^2), \tag{18}$$

where $T_k = \tau_k - \tau_{k-1}$ is the length of segment k, and $\widehat{\sigma}_k^2 = \frac{1}{T} \sum_{i=\tau_{k-1}}^{\tau_k} (Y_i - \overline{Y})^2$, is the empirical variance computed on that segment k and $\overline{Y}$ is the empirical mean of $Y_1, ..., Y_K$. When the true number K^* of segments is known, the sequence of change-points that minimizes this kind of contrast function has the property that, under extremely general conditions, for any $1 \leq k \leq K^* - 1$,

$$P(|\widehat{\tau}_k - \tau_k^*| > \delta) \longrightarrow 0, \text{ as } \delta \longrightarrow \infty, \ T \longrightarrow \infty, \tag{19}$$

where $\widehat{\tau}_k$ refers to the estimated and τ_k^* to the true segments of the breaks. This result holds for weakly and strongly dependent processes. When the number of change-points is unknown, it is estimated by minimizing a penalized version of the function $J(\tau,Y)$. For any sequence of change-points τ, let $pen(\tau)$ be a function of τ that increases with the number $K(\tau)$ of segments. Then, let $\{\widehat{\tau}\}$ be the sequence of change-points that minimizes

$$U(\tau) = J(\tau, \mathbf{Y}) + \beta_{\text{pen}}(\mathbf{Y}).$$

The penalty function is such as to avoid over- or under-segmentation. If β is a function of T that goes to 0 at an appropriate rate as $T \to \infty$, the estimated number of segments $K(\widehat{\tau}_k)$ converges in probability to K^* and condition (19) still holds.

The above estimation of multiple breaks and possibly the orders of the time series model is via a model selection approach of non-nested models. This method deals with the over-estimation of the number of breaks since it attaches a penalty term associated with the number of segments. The best combination of these values is then treated as an optimization of a desired contrast function. The literature uses various selection criteria for the multiple change-point problem. Early examples are Kitagawa and Akaike (1978) and Yao (1988) that use the Akaike and Bayesian Information Criteria (AIC and BIC), respectively, in the case of a change in the mean of an independent process. Chen and Gupta (1997) also consider the BIC criterion for locating the number of breaks in the variance of stock returns but still assume that the process is independent. More recently, for weakly dependent processes, Liu, Wu and Zidek (1997) modify the BIC by adding a larger penalty function and Bai and Perron (1998) consider criteria based on squared residuals. Lavielle and Moulines (2000) and Lavielle (1999) propose the penalized contrast function for selecting the sequence of change-points based on least squares (LS) and the optimal choice of the penalty function. This method can also be used for strongly dependent processes such as, for instance, financial time series that exhibit long memory. Consequently, this method can be viewed as detecting multiple breaks in a semiparametric model for financial time series.

For financial time series processes that follow a stochastic volatility process there are various additional challenging factors in multiple change-point detection. First, the models are nonlinear in nature which adds another level of difficulty on the optimization problem. Second, some volatility processes such as the stochastic volatility (SV) model do not have a closed form expression which makes estimation of multiple breaks for such models computationally challenging. Third, financial time series may exhibit strong dependence. Davis et al. (2005) present a method for detecting the optimal number and location of multiple change-points in SV, GARCH and other nonlinear processes based on the Minimum Description Length (MDL) criterion. Take, for instance, the multiple breaks GARCH model given by:

$$\begin{aligned} r_{tk}^2 &= \sigma_{tk}^2 u_t^2, \\ \sigma_{tk}^2 &= \alpha_{0,k} + a_{k,1}\sigma_{t-1,k}^2 + ... + a_{k,q_k}\sigma_{t-q_k,k}^2 + d_{k,1}r_{t-1,k}^2 + ... + d_{k,p_k}r_{t-p_k,k}^2, \\ &\quad \tau_{k-1} < t < \tau_k \end{aligned} \tag{20}$$

where u_t^2 is $NIID(0,1)$ and σ_{tk}^2 is a well-defined second-order stationary process. The unknown coefficients in this model include the parameter vector of GARCH coefficients as well as the orders of the model given by $\theta_k = (\alpha_{\mathbf{0},\mathbf{k}}, \alpha_k, \mathbf{d}_k, p_k, q_k)$. Given that the MDL principle can be expressed in terms of the log likelihood of the model, this method can also provide estimates of the orders of the model and its parameters in each segment.

The LS distance model selection change-point method in (18) is easier to implement (in terms of computational efficiency) and does not make any parametric and distributional assumptions. However, it does not reveal what is the actual change in the structure of the process, i.e., which parameters or orders of the time series are changing (e.g., drift, persistence in volatility), as opposed to the MDL method that applies to a parametric model but can disclose such information. Given that financial returns are heavy tailed, distances other than the LS may be of interest for the contrast function.

One of the challenges for these multiple change-point test methods is the optimization of the contrast function or criterion for the optimal combination of the number of segments, the length of the segments and possibly the parameters/orders of the time series model. Some of the optimization algorithms are, for instance, the genetic algorithm (Davis et al. (2005)) and Bellman and Roth (1969) and Guthery (1974) that uses least squares $O(T^2)$ operations.

A different approach to estimate the number and location of multiple breaks is based on the method of binary, sequential sample segmentation. Such methods were initially developed for the variance of an *i.i.d.* process (e.g., in Inclan and Tiao (1994), for the CUSUM of squares test) and further applied to the residuals of a GARCH model of emerging financial stock market indices (Aggarwal et al. (1999)) or the quadratic variation of the process (Andreou and Ghysels (2002)). This method addresses the issue of multiple change-points detection using a sample segmentation procedure to

sequentially or iteratively detect if a change-point is present in any subsample. This simple method can consistently estimate the number of breaks (e.g., Bai (1997), Inclan and Tiao (1994)). However, application especially for small samples must be cautioned by the fact that it might overestimate the number of breaks and their location may be wrong since the global change-point problem is translated into a sequence of local change-point detection problems.

4 Change–Point Tests in Returns and Volatility

In this section we discuss the applications of change-point tests for financial returns and volatility.

4.1 Tests based on empirical volatility processes

We consider empirical processes that obey a Functional Central Limit Theorem (FCLT) and we devote this subsection to the empirical processes and the conditions that need to hold to satisfy the FCLT. We are interested in strongly dependent time series, and in particular, stochastic volatility processes.

Since our main focus of interest is financial market volatility processes, we start from standard conditions in asset pricing theory. In particular, absence of arbitrage conditions and some mild regularity conditions imply that the log price of an asset must be a semi-martingale process (see e.g., Back (1991) for further details). Applying standard martingale theory, asset returns can then be uniquely decomposed into a predictable finite variation component, an infinite local martingale component and a compensated jump martingale component. This decomposition applies to both discrete and continuous time settings and is based on the usual filtration of past returns.

We will study data-driven processes related to volatility denoted by Y_t and defined by:

$$Y_T(\frac{sT}{T}) = \frac{1}{\sqrt{T}\widehat{\sigma}_T} \sum_{t=1}^{[sT]} Y_t, \tag{21}$$

where T is the sample size, s belongs $[0, 1]$ and $\widehat{\sigma}_T$ is the long-run variance estimator. For the purpose of FCLT arguments we can write:

$$Y_t = \widehat{\mu}_Y + \widehat{v}_t \qquad t = 1, \ldots, T, T+1, T+2, \ldots, \tag{22}$$

with $\widehat{v}_t = [Y_t - \widehat{\mu}_Y]$ and $\widehat{\mu}_Y = 1/T \sum_{t=1}^{T} Y_t \rightarrow_p \mu_Y$ as $T \rightarrow \infty$. Suitable regularity conditions will ensure that for various choices of Y_t, the partial sum

process Y_T in (21) will obey a FCLT. Moreover, various choices of Y_t imply different sample average limits μ_Y. It will be important to know the interpretation of μ_Y since the tests have local asymptotic power against alternatives that are characterized by perturbations of μ_Y.

Some of the empirical monitoring processes below relate more closely to the ARCH class of models, while other processes are more directly linked to SV-type models. We start with processes inspired by the ARCH literature.

4.1.1 The General ARCH class of models

A random sequence $\{Y_t, t \in \mathbf{T}\}$ satisfies the ARCH(∞) type equations if there exists a sequence of *i.i.d.* non-negative random variables $\{\xi_t, t \in \mathbf{T}\}$ such that:

$$Y_t = \sigma(\mathbf{Y}_{t-1}^0, \mathbf{b})\xi_t, \quad t \geq 1, \tag{23}$$

$$\sigma(\mathbf{Y}_{t-1}^0, \mathbf{b}) = b_0 + \sum_{j=1}^{\infty} b_j Y_{t-j}, \tag{24}$$

where $\mathbf{Y}_{t-1}^0 := (Y_{t-j}, j \geq 1)$, $b_0 \geq 0$, $b_j \geq 0$, $j = 1, 2, ...$ The model (23)-(24) specified in Robinson (1991) is general enough to include the following ARCH-type models. Let $Y_t := |r_t|^\delta$, $\xi_t := |\varepsilon_t|^\delta$ and $\sigma(Y_{t-1}^0, b_j) := \sigma_t^\delta$ such that:

$$|r_t|^\delta = \sigma_t^\delta |\varepsilon_t|^\delta, \quad \sigma_t^\delta = b_0 + \sum_{j=1}^{\infty} b_j |r_{t-j}|^\delta,$$

where $\delta > 0$ and ε_t are *i.i.d.* random variables with zero mean. In particular, for $\delta = 1$, we obtain Taylor's (1986) power or absolute value ARCH model

$$|r_t| = \sigma_t |\varepsilon_t|, \quad \sigma_t = b_0 + \sum_{j=1}^{\infty} b_j |r_{t-j}|, \tag{25}$$

and for $\delta = 2$, Engle's (1982) squared returns ARCH representation

$$r_t^2 = \sigma_t^2 \varepsilon_t^2, \quad \sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j r_{t-j}^2. \tag{26}$$

Moreover, Bollerslev's (1986) GARCH(p, q) model

$$r_t^2 = \sigma_t^2 u_t^2, \quad \sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} a_j \sigma_{t-j}^2 + \sum_{j=1}^{q} d_j r_{t-j}^2, \tag{27}$$

can be rewritten in the form of the general specification in (23)-(24) if $\sigma(Y_{t-1}^0, b_j) = \sigma_t^2$, $\xi_t = \varepsilon_t^2$, $Y_t = r_t^2$ and the coefficients $b_0 = \alpha_0(1 + \alpha_1 + \alpha_1^2 + ...) = \alpha_0/(1 - \alpha_1), b_j = a_1^{j-1} d_1, j = 1, 2, ...$

Before discussing the regularity conditions for the general model (23)-(24) to ensure a FCLT applies to partial sums of daily squared and/or absolute returns, it is first worth elaborating on the interpretation of μ_Y and therefore the power properties we expect. Equation (23) and the fact that ξ_t is *i.i.d.* implies that for empirical monitoring processes such as $Y_t = r_t^2$, then $\mu_Y = E\sigma_t^2 \times E\varepsilon_t^2$, whereas $\mu_Y = E\sigma_t \times E|\varepsilon_t|$ for empirical monitoring processes $Y = |r_t|$. Since by definition $\varepsilon_t \sim i.i.d.(0, 1)$, for both cases we have that $\mu_Y = E(\sigma_t^\delta)$ for $\delta = 1, 2$. Therefore, selecting either daily squared returns or absolute returns will result in tests that have power against various types of alternatives: (1) alternatives that change the mean volatility $E\sigma_t^2$ or $E\sigma_t$ and (2) alternatives that change the distribution of innovations, through $E|\varepsilon_t|$.

4.1.2 Regularity conditions for FCLT

If the above class of ARCH(∞) processes in (23)-(24) satisfies the following sufficient parameter and moment conditions:

$$E(\xi_0^2) < \infty, \qquad E(\xi_0^2) \sum_{j=1}^{\infty} b_j < 1, \tag{28}$$

then the following three properties hold for the stochastic process $\{Y_t\}$:

- is strictly and weakly stationary
- exhibits short memory in the sense that the covariance function is absolutely summable, $\sum_{t=-\infty}^{\infty} cov(Y_t, Y_0) < \infty$.
- satisfies the Functional Central Limit Theorem (FCLT).

Giraitis et al. (2000) prove that as $T \to \infty$

$$S_T(s) := T^{-1/2} \sum_{t=1}^{[sT]} (Y_t - E(Y_t)) \to \sigma W(s), \quad 0 \le s \le 1, \tag{29}$$

where $\sigma^2 = \sum_{t=-\infty}^{\infty} cov(Y_t, Y_0) < \infty$ and $\{W(\tau), 0 \le \tau \le 1\}$ is a standard Wiener process with zero mean and covariance $E(W(t)W(s)) = \min(t, s)$. It is interesting to note that the FCLT holds without having to impose any other memory restrictions on Y_t such as mixing conditions. The reason being that the condition in (28) implies not only (a) weak stationarity but also (b) short memory. In fact, the latter represents the key to the FCLT. Moreover, the autocorrelation function of Y_t depends directly on the existence of $E(\xi_0^2)$ (see for instance, He and Teräsvirta (1999)). In addition to the short memory structure of Y_t, Kokoszka and Leipus (2000) show that if the b_j in (23) decay exponentially then so does the covariance function. Similar results on the

behavior of the autocorrelation function of a GARCH(p, q) can be found in He and Teräsvirta (1999) and their results can be simplified to an ARCH model to yield the same condition as (28). Finally, under the sufficient condition (28) results (a) and (b) also imply (c). Note also that the FCLT holds without the Gaussianity assumption.

The ARCH-type models (26) and (27) can be considered in the context of the general specification (23)-(24) for which $\xi_t = f(\varepsilon_t)$ for some non-negative function f. Therefore, condition (28) can lead to corresponding conditions for these ARCH models. For instance, for Engle's ARCH(∞) model where $\xi_t = \varepsilon_t^2$, condition (28) becomes $E(\varepsilon_0^4) < \infty$ and $E(\varepsilon_0^4) \sum_{j=1}^{\infty} b_j < 1$ and for Taylor's ARCH model where $\xi_t = |\varepsilon_t|$ condition (28) becomes $E(\varepsilon_0^2) < \infty$ and $E(\varepsilon_0^2) \sum_{j=1}^{\infty} b_j < 1$. An alternative method for deriving the FCLT for a GARCH(p, q) process based on near-epoch dependence is found in Davidson (2002) who shows that a sufficient assumption is that ε_t is *i.i.d.* with finite fourth moment. One could consider the fourth moment existence condition imposed in the analysis to be restrictive.

The FCLT result for Y_t in (23) and (24) or equivalently, $Y_t = |r_t|^{\delta}, \delta = 1, 2$ in (26)-(27), provides the conditions for deriving the sequential CUSUM tests for the observed returns process in dynamic scale models. In contrast to linear dynamic regression models, the squared residuals of GARCH models do not satisfy the FCLT. Horváth et al. (2001) show that the partial sum of the ARCH(∞) squared residuals are asymptotically Gaussian, yet involve an asymptotic covariance structure that is a function of the conditional variance parameters and a function of the distribution of the innovations and moments of the ARCH. Consequently, boundary crossing probabilities can be computed for partial sums tests based on Y_t as opposed to the ARCH-type squared residuals.

4.2 Empirical processes and the SV class of models

In the previous subsection we dealt with daily returns. Here we consider processes Y_t based on intra-daily returns. Such processes are more closely related to the class of continuous time SV models. We discuss again under what regularity conditions partial sum processes appearing in (21) will obey a FCLT. Moreover, we also discuss the interpretation of μ_Y, and therefore the power properties the resulting tests are likely to have.

4.2.1 The General SV class of models

The purpose of using high frequency financial time series is to estimate more precisely and more directly volatility. There is now a substantial literature on the use of high frequency financial data, see, e.g., Andersen et al. (2003)

for a survey. In this section, we will examine two alternative processes for Y_t based on intra-daily returns, and to do so we start with the class of stochastic volatility models that is commonly used in financial economics to describe the behavior of asset returns. A typical continuous time SV model for log-prices $p(t)$ can be written as:

$$dp(t) = \nu(t)\,dt + \sigma(t)\,dW(t) + \kappa(t)\,dq(t) \tag{30}$$

where $dq(t)$ is a jump process with intensity $\lambda(t)$ size $\kappa(t)$. The process, $\nu(t)$ is a continuous locally bounded variation process, $\sigma(t)$ is a strictly positive and càdlàg stochastic volatility process and $W(t)$ is a Wiener process. Typically, the object of interest is to predict the increment of the quadratic variation over some horizon (typically daily), that is:

$$QV_{t,t+1} = \int_t^{t+1} \sigma^2(s)\,ds + \sum_{\{s\in[t,t+1]:dq(s)=1\}} \kappa^2(s). \tag{31}$$

The first component in equation (31) is sometimes written as:

$$\sigma^{[2]}_{t,t+1} = \int_t^{t+1} \sigma^2(s)\,ds. \tag{32}$$

Other measures have been studied as well, and it will be of particular interest to consider the alternative measure defined as:

$$\sigma^{[1]}_{t,t+1} = \int_t^{t+1} \sigma(s)\,ds. \tag{33}$$

The volatility measures appearing in equations (32) and (33) are not observable but can be estimated from data.

To proceed with estimation, we define the intra-daily returns. Recall that returns sampled at a daily frequency are denoted r_t. For the purpose of estimating volatility, we will also consider $r_{(m),t-j/m}$, the j^{th} discretely observed time series of continuously compounded returns with m observations per day (with the index $t-j/m$ referring to intra-daily observations). Hence, the unit interval for $r_{(1),t}$ is assumed to yield the daily return (with the subscript (1) typically omitted so that r_t will refer to daily returns). For example, when dealing with typical stock market data, we will use $m = 78$, corresponding to a five-minute sampling frequency. It is possible to consistently estimate $QV_{t,t+1}$ in (31) by summing squared intra-daily returns, yielding the realized variance:

$$RV^m_{t,t+1} = \sum_{j=1}^{m} \left(r_{(m),t+j/m}\right)^2. \tag{34}$$

When the sampling frequency increases, i.e. $m \to \infty$, then the realized variance converges uniformly in probability to the increment of the quadratic

variation, i.e.,

$$\lim_{m\to\infty} RV_{t,t+1}^{m} \to^{p} QV_{t,t+1}. \tag{35}$$

The second class of empirical processes related to volatility is known as Realized Absolute Value, or Power Variation, and is defined as:

$$PV_{t,t+1}^{m} = \sum_{j=1}^{m} \left| \left(r_{(m),t+j/m} \right) \right|, \tag{36}$$

where $\lim_{M\to\infty} PV_{t,t+1}^{m} \to^{p} \sigma_{t,t+1}^{[1]}$.

To complete the specification of the stochastic volatility process we need to augment equation (30) with a specification of the volatility dynamics. We start with a diffusion for $\sigma(t)$. Following Barndorff-Nielsen and Shephard (2001), we use a non-Gaussian Ornstein-Uhlenbeck (OU) process:

$$d\sigma(t) = -\delta\sigma(t)dt + dz(\delta t), \tag{37}$$

where $z(t)$ is a Lévy process with non-negative increments.

From the above diffusion one can compute, under suitable regularity conditions discussed later, population moments for stochastic processes $Y_t = RV_{t,t+1}^{m}$, $PV_{t,t+1}^{m}$. Such calculations appear, for example, in Barndorff-Nielsen and Shephard (2001) and Forsberg and Ghysels (2004). One can show that for $Y_t = RV_{t,t+1}^{m}$:

$$\mu_Y = E\sigma_{t,t+1}^{[2]} + E\int_{t}^{t+1} \kappa^2(t)\, dq(t),$$

whereas for $Y_t = PV_{t,t+1}^{m}$:

$$\mu_Y = E\sigma_{t,t+1}^{[1]}.$$

Therefore, we expect that tests based on $Y_t = RV_{t,t+1}^{m}$ will have power properties against alternatives characterized by changes in the volatility dynamics and changes in the distribution of jumps. In contrast, with $Y_t = PV_{t,t+1}^{m}$, we only expect to have power against alternatives driven by changes in the volatility process. Changes in the tail behavior, i.e., in the jump distribution, will not affect test statistics based on $Y_t = PV_{t,t+1}^{m}$. This distinction is important in practical applications (see also Woerner (2004), for further discussion). The impact of jumps and the choice of statistics and/or monitoring processes is still an open question.

4.2.2 Regularity conditions for FCLT

The regularity conditions for the application of FCLT to the partial sum process Y_T are in comparison to ARCH-type processes relatively straightforward when, for example, volatility follows a non-Gaussian OU process.

In particular, we can look upon the processes $PV^m_{t,t+1}$ and $RV^m_{t,t+1}$ as linear processes contaminated by measurement noise. For ARCH-type models, we dealt with strongly dependent processes, whereas here we consider processes directly related to the latent volatility dynamics and such processes are weakly dependent under the above assumptions. For example, De Jong (1997) and De Jong and Davidson (2002), obtain functional limit results for a broad class of serially dependent and heterogeneously distributed weakly dependent processes. The defining feature of such processes is that their normalized partial sums converge to processes having independent Gaussian increments, specifically, Brownian Motion in the case where the variances are uniformly bounded away from infinity and zero. Such results would apply to non-Gaussian OU processes as discussed above. Obviously, other volatility processes might require FCLT results of the type discussed in the previous section.

Let us consider again equation (37). This process yields an autocorrelation function $acf(\sigma, s) \equiv corr(\sigma(t), \sigma(t+s))$ equal to $acf(\sigma, s) = \exp(-\delta|s|)$. Using results from Barndorff-Nielsen and Shephard (2001), Mendoza (2004) obtains the following autocorrelation function:

$$acf(\sigma^{[1]}_{t,t+1}, s) = \frac{(1-e^{-\delta})^2 e^{-\delta(s-1)}}{2(e^{-\delta}-1+\delta)}, \tag{38}$$

whereas Konaris (2003) shows for the same process appearing in (37) that:

$$acf(\sigma(t)^2, s) = (1-\gamma)e^{-2\delta|s|} + \gamma e^{-\delta|s|} \tag{39}$$

$$acf(\sigma^{[2]}_{t,t+1}, s) = (1-\gamma)e^{-2\delta s}\left[\frac{(1-e^{-2\delta})^2}{4\delta^2}\right] + \gamma e^{-\delta s}\left[\frac{(1-e^{-\delta})^2}{\delta^2}\right], \tag{40}$$

where $\gamma = 2cov(\sigma(t), \sigma(t)^2)/var(\sigma^2(t))\widetilde{m}$ and $\widetilde{m}$ is the mean of $\sigma(t)$. Moreover,

$$cov(\sigma(t), \sigma(t)^2) = \kappa^\sigma_3 + 2\kappa^\sigma_2\kappa^\sigma_1 \tag{41}$$

$$var(\sigma(t)^2) = \kappa^\sigma_4 + 4\kappa^\sigma_3\kappa^\sigma_1 + 4\kappa^\sigma_2(\kappa^\sigma_1)^2 - (\kappa^\sigma_2)^2 \tag{42}$$

with κ^σ_i being the i^{th} order cumulant of the stationary process $\sigma(t)$. One can proceed by making a specific assumption about the marginal distribution of $\sigma(t)$.[2] Under regularity conditions discussed in detail in Barndorff-Nielsen and Shephard (2001) and Barndorff-Nielsen, Jacod and Shephard (2004), one ob-

[2] For example, one can assume that the law of $\sigma(t)$ is normal inverse Gaussian (henceforth NIG). This means that equation (37) is an $NIG-OU$ process. Assuming that the process (37) is a $NIG-OU$ puts restrictions on the so called background Lévy process $z(t)$. In particular, let the marginal be $NIG(\bar{\alpha}, \bar{\beta}, \mu, \delta)$ then the first four cumulants are (with $\rho = \beta/\alpha$):

$$\kappa^\sigma_1 = \mu + \frac{\delta\rho}{\sqrt{(1-\rho^2)}}, \quad \kappa^\sigma_2 = \frac{\delta^2}{\bar{\alpha}(1-\rho^2)^{3/2}},$$
$$\kappa^\sigma_3 = \frac{3\delta^3\rho}{\bar{\alpha}^2(1-\rho^2)^{5/2}}, \quad \kappa^\sigma_4 = \frac{3\delta^4(1+4\rho^2)}{\bar{\alpha}^3(1-\rho^2)^{7/2}}$$

tains stationary and ergodic processes $\sigma^{[i]}_{t,t+1}$, $i = 1,2$.[3] To establish the same properties for $Y_t = PV^m_{t,t+1}$, $RV^m_{t,t+1}$, we need to discuss the asymptotic distribution of the measurement error, that is, the difference between population processes $\sigma^{[i]}_{t,t+1}$, $i = 1,2$ and sampled processes $PV^m_{t,t+1}$ and $RV^m_{t,t+1}$.

Barndorff-Nielsen and Shephard (2001) show that in the absence of jumps, the error of realized variance is asymptotically (as $m \to \infty$), $RV_{t,t+1}$- $\sigma^{[2]}_{t,t+1}/\sqrt{2\sigma^{[4]}_{t,t+1}/3} \sim N(0,1)$, where $\sigma^{[4]}_{t,t+1} = \int_t^{t+1} \sigma(s)^4 ds$ is called the quarticity.[4] The error of the realized absolute variance can also be derived accordingly and yields $PV_{t+1,t}$ - $\sigma^{[1]}_{t+1,t} \sim N(0, 0.36338 RV_{t,t+1})$. While the asymptotic analysis for RV and PV were originally derived under different regularity conditions, recent work by Barndorff-Nielsen, Jacod and Shephard (2004) has provided a unified asymptotic treatment of both measures of volatility.

Given the above measurement process, it is clear that the sample process is stationary and ergodic under similar regularity conditions as the population processes and therefore the FCLT applies to both. The above results can also be broadened. When instantaneous volatility depends linearly on up through two autoregressive factors, Meddahi (2003) derives an ARMA representation of $RV^m_{t,t+1}$. The class of processes considered by Meddahi (2003) includes affine diffusions, GARCH diffusions, CEV models, as well as the OU-type processes appearing in equation (37). Consequently, with the high frequency data-based monitoring processes, we remain in the context of the linear processes considered by Kuan and Hornik (1995) and Leisch et al. (2000) since we monitor directly $RV^m_{t,t+1}$ as a weakly dependent ARMA process.

4.3 Tests based on parametric volatility models

This subsection discusses change-point tests which assume a specific parameterization of the financial returns and volatility process. For instance, Kulperger and Yu (2005) derive the properties of structural break tests based on the partial sums of residuals of GARCH models, whereas Berkes et al. (2004) present a likelihood-ratio (LR) based test for evaluating the stability of the GARCH parameters.

The properties of tests with unknown change-point based on the partial sums processes of residuals from parametric GARCH models for financial pro-

Forsberg and Ghysels (2004) take a different approach, which consists of selecting reasonable values for γ. Since the latter is equal to $2cov(\sigma(t), \sigma(t)^2)/var(\sigma^2(t))\widetilde{m}$ and $cov(\sigma(t), \sigma(t)^2)/var(\sigma^2(t)$ is the regression coefficient of a linear projection of $\sigma(t)$ onto $\sigma(t)^2$ one can select reasonable values for γ as well as $\widetilde{m}$ directly.

[3] We refrain from explicitly listing the regularity conditions as they are fairly mild, see also, Konaris (2003), Mendoza (2003), Forsberg and Ghysels (2004) and Woerner (2004).

[4] This result can be generalized to cases with jumps, see Forsberg and Ghysels (2004) for further discussion.

cesses can be found in Kulperger and Yu (2005). Here we focus on the CUSUM test for detecting changes in the mean and volatility of GARCH models. The asymptotic results of high moment partial sum processes of GARCH residuals in Kulperger and Yu (2005) can be extended to change-point tests in higher-order moments of GARCH residuals, such as, for instance, moments that relate to the asymmetry and tails of financial processes. This residual-based CUSUM test involves the same conditions on a GARCH model as other tests (e.g., Kokoszka and Leipus (2000), and Horváth et al. (2001)) which are essentially fourth order stationarity and $\sqrt{n}$ consistency of the volatility estimator. Note that no specific distributional assumptions are imposed other than the GARCH errors being $i.i.d.(0,1)$. However, under the assumption of a symmetric distribution for the innovations, the asymptotic distribution of the standardized $kth-$order moment of the residual centered partial sum process of a GARCH given by

$$T_n^{(k)}(s) = \sum_{t=1}^{[sT]} (\widehat{u}_t - \overline{\widehat{u}})^k, 0 \leq s \leq 1,$$

is a Brownian Bridge with no nuisance parameters. On the other hand, if no symmetry assumption is imposed, then for $k > 3$ the asymptotic Gaussian process depends on the moment of the innovation distribution and cannot be identified with a specific classic process such as a Brownian Motion or Brownian Bridge. Under the null hypothesis of no structural breaks, the GARCH(p,q) model yields the error $u_t = r_t/\sigma_t$ where $\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} a_j \sigma_{t-j}^2 + \sum_{j=1}^{q} d_j r_{t-j}^2$. Under the alternative there may be a break in the conditional mean or the conditional variance of the GARCH given by

$$r_t = \sigma_t u_t + \mu, \quad \sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} a_j \sigma_{t-j}^2 + \sum_{j=1}^{q} d_j \left(r_{t-j} - \mu\right)^2, \tag{43}$$

$\mu \neq 0 \quad t = [\tau T] + 1, ..., T$, and

$$r_t = \sigma_t u_t, \ \sigma_t^2 = \begin{cases} \alpha_0 + \sum_{j=1}^{p} a_j \sigma_{t-j}^2 + \sum_{j=1}^{q} d_j r_{t-j}^2 & \text{if } t = 0, ..., [\tau T] \\ \alpha_0' + \sum_{j=1}^{p} a_j' \sigma_{t-j}^2 + \sum_{j=1}^{q} d_j' r_{t-j}^2 & \text{if } t = [\tau T] + 1, ..., T \end{cases} \tag{44}$$

respectively. The residual CUSUM test for detecting breaks in the conditional mean is:

$$\begin{aligned} CUSUM^{(1)} &= \max_{1 \leq i \leq n} \frac{\left|\sum_{t=1}^{i} (\widehat{u}_t - i\overline{\widehat{u}})\right|}{\widehat{\sigma}_{(n)}^2 \sqrt{n}} = \max_{1 \leq i \leq n} \frac{1}{\sqrt{n}} \left|\sum_{t=1}^{i} (\widehat{u}_t - i\overline{\widehat{u}})\right| \\ &\rightarrow \sup |B_0(\tau)| \end{aligned} \tag{45}$$

since for a GARCH model, $\widehat{\sigma}^2_{(n)}$ is an estimator of the $var(u_t) = 1$, by definition. Similarly, the squared residual CUSUM test for detecting breaks in the conditional variance is:

$$CUSUM^{(2)} = \max_{1\leq i\leq n} \frac{\left|\sum_{t=1}^{i} \widehat{u}_t^2 - i\sum_{t=1}^{n} \widehat{u}_t^2/n\right|}{\widehat{\nu}_2\sqrt{n}}$$

$$= \max_{1\leq i\leq n} \frac{\left|\sum_{t=1}^{i} \left(\widehat{u}_t - \overline{\widehat{u}}\right)^2 - i\widehat{\sigma}^2_{(n)}\right|}{\widehat{\nu}_2\sqrt{n}} \to \sup|B_0(\tau)|, \quad (46)$$

where $\widehat{\nu}_2 = \frac{1}{n}\sum_{t=1}^{i}\left(\left(\widehat{u}_t - \overline{\widehat{u}}\right)^2 - \widehat{\sigma}^2_{(n)}\right)^2$ is the estimator of $\nu_2 = E(u_0^2 - E(u_0^2))^2$. Given the asymptotic properties of the residual partial sums processes, it is possible to obtain the asymptotic distribution of other types of test statistics similar to the Fluctuation test and the Page tests. Compared to the CUSUM tests for detecting breaks in $Y_t = |r_t|$ or r_t^2, discussed in the previous sections, the residual-based CUSUM tests for detecting breaks in the mean and variance of a GARCH model do not involve the estimation of a long-run matrix using Heteroskadastic and Autocorrelation Consistent (HAC) estimators. Moreover, the results in Kulperger and Yu (2005) show that under certain cases these tests have better finite sample properties than the returns based CUSUM tests, e.g., in Kokoszka and Leipus (2000). It is also worth noting that the Chen et al. (2005) test discussed in section 3.2 is also a CUSUM-based residual test which is, however, based on the nonparametric estimation of more general specifications.

Given that financial processes exhibit heavy tails, Andreou and Werker (2005) present the asymptotic distribution of a CUSUM test based on the ranks of the residuals from a GARCH model for detecting change-points. The statistic does not involve any nuisance parameters and also converges to the same asymptotic distribution. Hence, it is not only robust to alternative distributional assumptions but may exhibit better power in detecting breaks in heavy-tailed financial processes. In addition, it does not involve the standardization by a long-run variance estimator (compared to the CUSUM tests for the observed returns processes).

Unlike the above parametric method that relies on the residuals of the GARCH model, the method proposed by Berkes et al. (2004) is based on quasi-likelihood scores and can be used to evaluate which of the parameters of a GARCH(p, q) has a change-point. In the general setup, the observed financial process $r_1, ..., r_n$ may follow a GARCH(p, q) model with d parameters. Denote by ω a generic element in the parameter space and by $\ell_i(\omega)$ the conditional quasi-likelihood of r_i given by $r_{i-1}, ..., r_1$, so that the quasi-likelihood function is $L_m(\omega) = \sum_{1\leq i\leq m} \ell_i(\omega)$. For time series models, $\ell_i(\omega)$ can not be computed exactly because of the dependence on the unobserved $d-$dimensional row vector of partial derivatives with respect to the model

parameters. Consider the matrix

$$\widehat{\mathbf{D}}_n = \frac{1}{n} \sum_{1<i\leq n} \left(\widehat{\ell}_i'(\widehat{\theta}_n)\right)^T \left(\ell_i(\widehat{\theta}_n)\right),$$

where $\widehat{\theta}_n$ is the quasi maximum likelihood parameter estimate. The d-dimensional process $\mathbf{G_m} = \sum_{1<i\leq n}(\widehat{e}_i'(\widehat{\theta}_n)\widehat{\mathbf{D}}_\mathbf{n}^{-1/2}$ can form the basis of test statistics based on appropriate approximations for $\widehat{\ell}_i'(\widehat{\theta}_n)$. Berkes et al. (2004) derive a sequential likelihood ratio test for monitoring the parameters of the GARCH model which is more informative than any sequential CUSUM test performed on the observed returns process or residual transformations.

4.4 Change–point tests in long memory

It is widely documented that various measures of stock return volatility (e.g., squared and absolute returns) exhibit properties similar to those of a long-memory process (e.g., Ding et al. (1993), Granger and Ding (1995) and Lobato and Savin (1998)). More recent evidence supports the view that stock market volatility may be better characterized by a short-memory process affected by occasional level shifts found, for instance, in Mikosch and Stărică (2004), Perron and Qu (2004) and Granger and Stărică (2005). This pattern, found in daily SP500 absolute returns, is very close to what is expected with a short-memory process with level shifts. The interplay between structural breaks and long memory demonstrates that by accounting for structural breaks, the estimates of the long-memory parameters in stock return volatility within regimes are reduced (e.g., Granger and Hyung (2004)). Moreover, superior forecasts of exchange rate returns can be obtained in longer horizons by modeling both long memory and structural breaks (Beltratti and Morana (2006)). In addition, it has been documented that short-memory processes with level shifts will exhibit properties that make standard tools conclude that long memory is present (e.g., Diebold and Inoue (2001), Engle and Smith (1999), Granger and Hyung (2004), Lobato and Savin (1998), Mikosch and Stărică (2004)). Hence, it is empirically difficult to discriminate a long-memory process from a weakly dependent process with some form of nonstationarity such as regime switching or structural breaks in the mean or volatility. Furthermore, Giraitis et al. (2001) provide analytical results to the above debate by showing that a structural change of a constant magnitude in linear and ARCH models which does not decrease with the sample size, will be picked up as long memory with probability approaching one as the sample size T, tends to infinity.

A recent test proposed by Berkes et al. (2006) might shed more light in the aforementioned empirical debate as a method to discriminate between a long-

memory dependent process and a weakly dependent process with changes in the mean or volatility of financial time series. In its simplest form, the test assumes that under the null hypothesis the process is weakly dependent with one unknown break in the mean and under the alternative it is a process with long-memory. The test procedure is based on the following: The CUSUM statistic is computed, as defined in previous sections, given by,

$$S_T^{CS} = (T\widehat{\sigma_T}^2)^{-1/2} \max \left| \sum_{1\le t\le k} Y_t - \frac{k}{T} \sum_{1\le t\le T} Y_t \right| \tag{47}$$

where $\widehat{\sigma_T}^2$ is the long variance estimator of the sample mean of Y_t. For financial time series, Y_t may again represent squared or absolute returns given the empirical evidence of long memory. The value of the statistic at $\max |S_n^{CS}|$ is used to segment the sample at a point $\widehat{\tau}_1 = \max |S_n^{CS}|$ whether there is a structural break or not. Then the CUSUM statistic is computed in the two segmented samples up to $\widehat{\tau}_1$ given by $S_{T,1}^{CS}$ and from $\widehat{\tau}_1 + 1$ to the end of the sample given $S_{T,2}^{CS}$. Under the null hypothesis, the resulting asymptotics of the statistic obtained from (47) in each sub-sample is given by

$$M_1 = \max \left[S_{T,1}^{CS}, S_{T,2}^{CS} \right] \to \max \left[\sup_{0\le t\le 1} |B^{(1)}(t)|, \sup_{0\le t\le 1} |B^{(2)}(t)| \right], \tag{48}$$

see Kiefer (1959). Under the alternative, the test statistic diverges to infinity. Note that this test is based on the almost sure asymptotics for the long-run Bartlett variance estimator σ_T^2.

This test can be easily extended to examine the null hypothesis of a weakly dependent process with k multiple change-points versus the long-memory alternative, using the sequential, binary, sample segmentation approach (discussed at the end of section 3.3). The asymptotic distribution of the test statistic now generalizes to the $k+1$ analogue of (48) that involves CUSUM statistics in $k+1$ regimes and the null hypothesis is examined sequentially at each sample segmentation stage.

Related tests for multiple structural changes in a long memory process based on a least-squares model selection approach can be found in Lavielle and Moulines (2000). For a test in the long memory parameter based on the maximal difference across potential break dates of appropriately weighted sums of autocovariances, see Beran and Terrin (1996).

Some popular tests such as Hurst's rescaled range type statistic for long-memory are also related to tests for structural breaks. The weakness of these tests is that they can not discriminate between long-range dependence and weak-dependence with structural change, compared to the aforementioned Berkes et al. (2006) test. Giraitis et al. (2003) also propose the rescaled variance test V/S, based on the sample variance of the partial sum process:

$$V/S(q) = \frac{\widehat{var}(S_1, ..., S_T)}{T\widehat{\sigma}^2(q)} = \frac{1}{T^2\widehat{\sigma}^2(q)} \left[\sum_{k=1}^{T} S_k^2 - \frac{1}{T} \left(\sum_{k=1}^{T} S_k \right)^2 \right] \quad (49)$$

where $\widehat{\sigma}^2(q)$ is the long run variance estimator. They find that it is more sensitive to changes in the variance and would have higher power than the rescaled range statistic against long memory in the squares.

In the above tests for long-memory in volatility models the bandwidth parameter q of the long-run variance estimator plays a special role. The question on the optimal q is still open and the properties of the above tests for detecting structural breaks in the long memory in view of the role of q need further investigation. Related to this is the investigation of the properties of the above tests with other long-run volatility estimators that deal with long memory such as, for instance, those proposed in Robinson (2005) and Abadir et al. (2006).

4.5 Change–point in the distribution

This section discusses tests for detecting changes in the distribution function of financial returns. The stylized fact of non-Normality in the asset returns is well documented in the empirical finance literature. More precisely, the properties of heavy tails, asymmetries and a large class of alternative distributions have been fitted to asset returns with no empirical consensus regarding a benchmark distribution.

Nonparametric change-point tests in the distribution of a strongly mixing process are proposed, for instance, in Inoue (2001) and Lavielle (1999). Such nonparametric tests are motivated by the robustness against misspecification as compared to analogous parametric and semiparametric tests, e.g., in Horváth et al. (2001), which nevertheless have more power under the assumption of a well-specified model.

The tests proposed in Inoue (2001) are nonparametric in the sense that they do not specify a distribution nor a specific parametric model for the asset returns process and are based on the difference between empirical distribution functions (edf). These tests have at least two advantages compared to nonparametric density estimators: The edf test convergence rate is always $\sqrt{T}$, and it does not suffer from the dimensionality curse. In contrast, it is well known that tests based on nonparametric density estimators suffer from the curse of dimensionality and have a slower than $\sqrt{T}$ convergence rate which may not have power against $\sqrt{T}$ local alternatives. Two additional features of these edf based change-point tests which are useful for financial time series are their robustness to heavy tails and to nonlinear dependence as well as their robustness to the inexistence of the unconditional fourth moment (Inoue (2001)). This nonparametric edf based test allows dependence and

consequently, its limiting null distribution depends on a nuisance parameter which is derived using bootstrap methods. This test is based on the limiting process of a simulated sequential empirical process. The simulated-based test has power against local alternatives and is consistent against multiple breaks. However, this nonparametric edf-based test largely depends on the size of the block bootstrap, and the asymptotic behavior of the selected block length under the alternative hypothesis is still unexplored.

A complementary test to detect structural changes in the distribution is based on an edf process of the residuals of volatility models for financial returns. Horváth, Kokoszka and Teyssière (2001) show that unlike the residuals of ARMA processes (e.g., Bai (1994)), the residuals of the ARCH models yield sequential empirical processes that do not behave like asymptotically independent random variables. In particular, they show that the asymptotic distribution involves, among others, a term depending on the unknown parameters of the model. For ARCH models the detection of changes in the distribution function of unobserved innovations yields sequential edf tests that lead to asymptotically distribution free statistics.

The above edf-based tests are based on the observed returns process or the residuals of a model for returns. Although the residual-based edf test is relatively easier to implement given its nuisance-parameter free limiting distribution, it however depends on the crucial assumption of a correctly specified ARCH-type model. These two edf tests will have different properties whether or not the correct model specification is assumed. It is useful if the two tests are viewed in a complementary approach. In view of the alternative distribution families proposed for financial time series and the alternative model specifications, if the correct parameterization is unknown then the nonparametric edf based test can serve as a useful pretest of the null hypothesis of distributional homogeneity. However, under the correct specification, the parametric edf tests would have more power. Moreover, the parametric edf tests or any of the other parametric change-point tests discussed here would be more informative as to the source of the structural change.

Another fundamental difference between the above two tests is that the residual-based edf test examines the homogeneity in the conditional distribution of returns whereas the returns-based edf test assesses the homogeneity of the marginal distribution of returns. An alternative nonparametric test for the stability of the marginal distribution of strongly dependent *and* strongly mixing processes that also aims to detect unknown multiple breaks is based on minimizing a penalized contrast function, proposed by Lavielle (1999). It is assumed that the distribution of such processes depends on a parameter θ that changes abruptly at some points. When the number of change-points is known, their configuration is estimated by minimizing a contrast function. It is shown, under mild assumptions, that, if the minimum contrast estimate of θ, computed in any segment of the true configuration of change-points, is consistent, then the change-points are consistently estimated. Moreover, the estimated parameter vector of θ_n also converges to the true vector of param-

eters θ^*. When the number of change-points is known, the convergence rate of $\|\widehat{\tau}_n - \tau\| \to O_p(n^{-1})$ does not depend on the covariance structure of the process, whereas the convergence of $\widehat{\theta}_n$ depends on this covariance.

5 Conclusions

This review deals with a part of the literature on structural breaks tests for financial time series. A review of the related literature on structural breaks in measures of co-dependence of financial time series and in asset pricing models is found in Andreou and Ghysels (2006a). In concluding, we point to some important questions that still remain unaddressed and some interesting issues that require further progress in this area.

Further research of tests for unknown change-points in systems of equations such as multivariate volatility models with ARCH or long memory type effects is largely unexplored. This is especially true for endogenous breaks tests in copulae models which form a parsimonious way of capturing a multivariate process of financial returns, volatility and other forms of non-linearities. There is some work on change-point tests in bivariate models of conditional volatility and co-dependence (e.g., Andreou and Ghysels (2003)). Generalizing change-point tests in multivariate systems to multiple breaks that can be detected in different equations and may affect a subset of the variables is still a challenge. In addition, there is less research on structural break tests for continuous time stochastic volatility models when the change-point is unknown.

Related to all structural change tests is the issue of robustness. Given the evidence of non-linear, short and long memory, heavy-tailed, asymmetric mechanism of financial asset returns, it is useful that the estimated change-points in empirical studies are robust towards some of these attributes and are not the artifact of misspecification. Some recent research supports the view that financial stock returns exhibit weak dependence and structural breaks as opposed to strong dependence. Recent work by Berkes et al. (2006) sheds more light in the memory and structural breaks debate in the mean of time series processes which would be interesting to extend to the volatility of financial processes that exhibit second-order dependence and/or long memory. Further analysis as to the long memory versus short memory and breaks dichotomy, especially in view of the plausible multiple change-points question in long samples of financial returns and the long memory in volatility based on high-frequency processes, requires further investigation.

Another direction towards this issue involves analytical asymptotic local power results of change-point tests with varying sampling frequencies which, for financial time series (unlike linear time series), must take into account the different persistence and tail behavior, e.g., Andreou and Ghysels (2006). Since the sampling frequency is often a choice variable for financial time series

and since there is no measurement error or any high cost in sampling more frequently, the sequential change-point tests for certain financial variables have various advantages and can also matter for the power of the tests.

Finally, it may be worth thinking further the economic significance of structural breaks in the financial models and a mechanism other than an exogenous determination of capturing reoccurring breaks. In most of the aforementioned papers, structural breaks in financial processes are associated with external events to the stochastic financial process. Recent research attempts to endogenize breaks by incorporating them in a Bayesian estimation and prediction procedure that allows for such structural changes (e.g., Pesaran et al. (2006)) or allows time variation in the model parameters of volatility that is assumed to be only locally homogeneous (e.g., Dalhaus and Rao (2006), Mercurio and Spokoiny (2004)). The relationship and empirical performance of time varying volatility models with multiple breaks ARCH models as well as the former's consequences for long memory and tail behavior are also interesting areas of future research.

References

Abadir, K., Distasio, W. and Giraitis, L. (2006): Two estimators of the long-run variance. *Working paper, University of York*.

Aggarwal, R., Inclan, C. and Leal, R. (1999): Volatility in emerging stock markets. *The Journal of Financial and Quantitative Analysis* **34**, 33–55.

Andersen, T. G. (1994): Stochastic autoregressive volatility: A framework for volatility modeling. *Mathematical Finance* **4**, 75–102.

Andersen, T. G., Bollerslev, T. and Diebold, F. X. (2003): Parametric and nonparametric volatility measurement. *In: Hansen, L.P. and Aït-Sahalia, Y. (Eds.): Handbook of Financial Econometrics*. North-Holland, Amsterdam. Forthcoming.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2001): The distribution of exchange rate volatility. *Journal of American Statistical Association* **96**, 42–55.

Andreou, E. and Ghysels, E. (2002): Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics* **17**, 579–600.

Andreou, E. and Ghysels, E. (2003): Tests for breaks in the conditional co-movements of asset returns. *Statistica Sinica* **13**, 1045–1073.

Andreou, E. and Ghysels, E. (2005): Quality control for financial risk management. *Discussion Paper, UNC*.

Andreou, E. and Ghysels, E. (2006a): Monitoring disruptions in financial markets.- *Journal of Econometrics* **135**, 77–124.

Andreou, E. and Ghysels, E. (2006b): Structural Breaks in Financial Time Series. *SSRN Working Paper* `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=935971`

Andreou, E. and Ghysels, E. (2007): Quality control for structural credit risk models. *Journal of Econometrics* forthcoming.

Andreou, E. and Werker, B.J.M. (2005): An Alternative Asymptotic Analysis of Residual-Based Statistics. *Working paper, Tilburg University*.

Back, K. (1991): Asset pricing for general processes. *Journal of Mathematical Economics* **20**, 305–371.

Bai, J. (1994): Weak Convergence of Sequential Empirical Processes of Residuals in ARMA Models. *Annals of Statistics* **22**, 2051–2061.

Bai, J. (1997): Estimating multiple breaks one at a time. *Econometric Theory* **13**, 315–352.
Bai, J. and Perron, P. (1998): Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47–78.
Barndorff-Nielsen, O. and Shephard, N. (2001): Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B* **64**, 253–280.
Barndorff-Nielsen, O. and Shephard, N. (2004): Limit theorems for bipower variation in financial econometrics. *Discussion Paper, Aarhus and Oxford University.*
Bates, D.J. (2000): Post-'87 crash fears in the s&p500 futures option markets. *Journal of Econometrics* **94**, 181–238.
Bekaert, G., Harvey, C.R. and Lumsdaine, R.L. (2002): Dating the integration of world equity markets. *Journal of Financial Economics* **65**, 203–247.
Bellman, R. and Roth, R. (1969): Curve fitting by segmented straight lines. *Journal of the American Statistical Association* **64**, 1079–1084.
Beran, J. and Terrin, N. (1996): Testing for a change in the long-memory parameter. *Biometrika* **83**, 627–638.
Berkes, I., Gombay, E., Horváth, L. and Kokoszka, P. (2004): Sequential change-point detection in GARCH(p,q) models. *Econometric Theory* **20**, 1140–1167.
Berkes, I., Horváth, L., Kokoszka, P. and Shao, Q. (2006): On discriminating between long-range dependence and changes in mean. *The Annals of Statistics* **34**, 1140–1165.
Beltratti, A. and Morana, C. (2006): Breaks and persistency: macroeconomic causes of stock market volatility. *Journal of Econometrics* **131**, 151–177.
Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.
Bollerslev, T., Engle, R.F. and Nelson, D.B. (1994): ARCH models. *In: Engle, R.F., McFadden, D. (Eds.): Handbook of Econometrics, Vol. IV*, 2959–3038. North-Holland, Amsterdam.
Brown, R.L., Durbin, J. and Evans, J.M. (1975): Techniques for testing the constancy of regression relationships over time. *Journal of Royal Statistical Society* **B**, 149–192.
Carrasco, M. and Chen, X. (2002): Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17–39.
Chang-Jin, K., Morley, J.C. and Nelson, C.R. (2005): The Structural Break in the Equity Premium. *Journal of Business and Economic Statistics* **23**, 181–191.
Chen, G., Choi, Y.K. and Zhou, Y. (2005): Nonparametric estimation of structural change points in volatility models for time series. *Journal of Econometrics* **126**, 79–114.
Chen, J. and Gupta, A.K. (1997): Testing and locating variance change points with application to stock prices. *Journal of the American Statistical Association* **92**, 739–747.
Chernov, M., Gallant, A.R., Ghysels, E. and Tauchen, G. (2003): Alternative models for stock price dynamics. *Journal of Econometrics* **116**, 225–257.
Chu, C.-S. (1995): Detecting parameter shift in GARCH models. *Econometric Reviews* **14**, 241–266.
Dahlhaus, R. and Rao, S.S. (2006): Statistical inference for time-varying ARCH process. *The Annals of Statistics* **34**, 1075–1114.
Davidson, J. (2002): Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes. *Journal of Econometrics* **106**, 243–269.
Davis, R.A. and Mikosch, T. (1998): The sample ACF of Heavy-Tailed Stationary Processes with Applications to ARCH. *The Annals Statistics* **26**, 2049–2080.
Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2005): Break detection for a class of nonlinear time series model. *Working Paper, Colorado State University.*
De Jong, R.M. (1997): Central limit theorems for dependent heterogeneous random variables. *Econometric Theory* **13**, 353–367.
De Jong, R.M. and Davidson, J. (2002): The functional central limit theorem and weak convergence to stochastic integrals I: weakly dependent processes. *Econometric Theory* **16**, 621–642.

Diebold, F.X. (1986): Modeling the Persistence of Conditional Variances: A comment. *Econometric Reviews* **5**, 51–56.

Diebold, F.X. and Inoue, A. (2001): Long memory and regime shifts. *Journal of Econometrics* **105**, 131–159.

Ding, Z., Granger, C.W.J. and Engle, R.F. (1993): A long-memory properties of stock market returns and a new model. *Journal of Empirical Finance* **2**, 83–106.

Drost, F.C and Nijman, T. (1993): Temporal Aggregation of GARCH Processes. *Econometrica* **61**, 727–909.

Engle, R.F. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987–1008.

Engle, R. and Smith, A. (1999): Stochastic permanent beaks. *The Review of Economics and Statistics* **81**, 553–574.

Forsberg, L. and Ghysels, E. (2004): Why do absolute returns predict volatility so well? *Discussion Paper, UNC.*

Foster, D. and Nelson, D. (1996): Continuous record asymptotics for rolling sample estimators. *Econometrica* **64**, 139–174.

Garcia, R. and Ghysels, E. (1998): Structural change and asset pricing in emerging markets. *Money and Finance* **17**, 455–473.

Genon-Catalot, V., Jeantheau, T. and Lared, C. (2000): Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* **6**, 1051–1079.

Ghysels, E., Harvey, A. and Renault, E. (1996): Stochastic Volatility. *In Maddala, G.S. (Ed.): Handbook of statistics, Statistical Methods in Finance* **14**. North, Holland, Amsterdam.

Giraitis, L., Kokoszka, P. and Leipus, R. (2000): Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory* **16**, 3–22.

Giraitis, L., Kokoszka, P., Leipus, R. and Teyssière, G. (2003): Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics* **112**, 265–294.

Granger, C.W.J. and Ding, Z. (1996): Varieties of long memory models. *Journal of Econometrics* **73**, 61–77.

Granger, C.W.J. and Hyung, N. (1999): Occasional structural breaks and long memory. *Working Paper.*

Granger C. and Stărică, C. (2005): Non-stationarities in stock returns. *Review of Economics and Statistics* **87**, 503–522.

Guthery, S.B. (1974): Partition Regression. *Journal of the American Statistical Association* **69**, 945–947.

He, C. and Teräsvirta, T. (1999): Properties of moments of a family of GARCH processes. *Journal of Econometrics* **2**, 173–192.

Hillebrand, E. (2005): Neglecting parameter changes in GARCH models. *Journal of Econometrics* **129**, 121–138.

Horváth, L., Kokoszka, P. and Teyssière, G. (2001): Empirical process of the squared residuals of an ARCH sequence. *Annals of Statistics* **29**, 445–469.

Horváth L., Kokoszka, P. and Zhang, A. (2006): Monitoring constancy of variance in conditionally heteroskedastic time series. *Econometric Theory* **22**, 373–402.

Inclan, C. and Tiao, G.C. (1994): Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association* **89**, 913–923.

Inoue, A. (2001): Testing for distributional change in time series. *Econometric Theory* **17**, 156–187.

Kiefer J. (1959): K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-V. Mises Tests. *The Annals of Mathematical Statistics* **30**, 420–447.

Kitagawa, G. and Akaike, H. (1978): A procedure for the modelling of non-stationary time series. *Annals of the Institute of Statistical Mathematics Part B* **30**, 351–363.

Kokoszka, P. and Leipus, R. (2000): Change-point estimation in ARCH models, *Bernoulli* **6**, 513–539.

Konaris, G. (2003): Derivative pricing under non-Gaussian stochastic volatility. *D. Phil. Thesis, Department of Economics, Oxford University.*

Kulperger R. and Yu, H. (2005): High moment partial sum processes of residuals in GARCH models and their applications. *Annals of Statistics* **33**, 2395–2422.

Kuan, C.M. and Hornik, K. (1995): The generalized fluctuation test: a unifying view. *Econometric Reviews* **14**, 135–161.

Lamoureux, C. and Lastrapes, W. (1990): Persistence in variance, structural GARCH model. *Journal of Business and Economic Statistics* **8**, 225–234.

Lavielle, M. (1999): Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications* **83**, 79–102.

Lavielle, M. and Moulines, E. (2000): Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis* **21**, 33–59.

Leisch, F., Hornik, K. and Kuan, C.M. (2000): Monitoring structural changes with the generalized fluctuation test. *Econometric Theory* **16**, 835–854.

Liu, J., Wu, S. and Zidek, J.V. (1997): On segmented multivariate regressions. *Statistica Sinica* **7**, 497–525.

Lobato, I.N. and Savin, N.E. (1998): Real and spurious long memory properties of stock market data. *Journal of Business and Economic Statistics* **16**, 261–283.

Lundbergh, S. and Teräsvirta, T. (1998): Evaluating GARCH models. *Technical Report* ***292****. Stockholm School of Economics.*

Meddahi, N. (2003): ARMA representation of integrated and realized variances. *Discussion Paper CIRANO.*

Meddahi, N. and Renault, E. (2004): Temporal aggregation of volatility models. *Journal of Econometrics* **119**, 355–379.

Mendoza, C. (2004): Realized variance and realized absolute variation. *Department of Statistics, Oxford University.*

Mercurio, D. and Spokoiny, V. (2004): Statistical Inference for time-inhomogeneous volatility models. *Annals of Statistics* **32**, 577–602.

Mikosch, T. and Stărică, C. (2004): Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *The Review of Economics and Statistics* **86**, 378–390.

Page, E. (1954): Continuous inspection schemes. *Biometrika* **41**, 100–115.

Page, E. S. (1955): A test for a change in a parameter occurring at an unknown point. *Biometrika* **42**, 523–527.

Pastor, L. and Stambaugh, R.F. (2001): Equity premium and structural breaks. *Journal of Finance* **56**, 1207–1239.

Perron, P. and Qu, Z. (2004): An analytical evaluation of the log-periodogram estimate in the presence of level shifts and its implications for stock return volatility. *Manuscript, Department of Economics, Boston University.*

Pesaran, H., Pettenuzzo, D. and Timmermann, A. (2006): Learning, structural instability and present value calculations. *Discussion Paper.*

Pesaran, M.H. and Timmermann, A. (2004): How costly is to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting* **20**, 411–425.

Pettenuzzo D. and Timmerman, A. (2005): Predictability of Stock Returns and Asset Allocation under Structural Breaks. *UCSD Working Paper.*

Phillip, W. and Stout, W. (1975): Almost sure invariance principles for partial sums of weakly dependent random variables. *Memoirs of the American Statistical Society* **161**.

Robinson P. (1991): Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* **47**, 67–84.

Robinson P. (2005): Robust Covariance Matrix Estimation: 'HAC' Estimates with Long Memory/Antipersistence Correction. *Econometric Theory* **21**, 171–180.

Shephard, N. (2005): Stochastic volatility: selected readings. *Edited volume, Oxford University Press* forthcoming.

Taylor, S. (1986): Modelling financial time series. *Wiley, New York.*

White, H. (1984): Asymptotic theory for Econometricians. *Academic Press.*

Woerner, J. (2004): Variational sums and power variation: a unifying approach to model selection and estimation in semimartingale models. *Statistics and Decision* forthcoming.

Yao, Y.C. (1988): Estimating the number of change-points via Schwartz criterion. *Statistics and Probability Letters* **6**, 181–189.

An Introduction to Regime Switching Time Series Models

Theis Lange and Anders Rahbek

Abstract A survey is given on regime switching in econometric time series modelling. Numerous references to applied as well as methodological literature are presented. A distinction between observation switching (OS) and Markov switching (MS) models is suggested, where in OS models, the switching probabilities depend on functions of lagged observations. In contrast, in MS models the switching is a latent unobserved exogenous process. With an emphasis on OS and MS ARCH and cointegrated models, stationarity and ergodicity properties are discussed as well as likelihood-based estimation, asymptotic theory and hypothesis testing.

1 Introduction

This survey considers regime switching time series models, which are models that allow parameters of the conditional mean and variance to vary according to some finite-valued stochastic process with states or regimes s_t, $s_t \in \mathbb{S} = \{1, ..., r\}$. The regime changes reflect, or aim at capturing, changes in the underlying financial and economic mechanism through the observed time period. For instance, as was argued in Bec and Rahbek (2004) and the references therein, term-structure and exchange rate data seem to exhibit epochs of both non-stationary and also mean-adjusting behavior. Likewise, as put forward e.g. in Lamoureux and Lastrapes (1990), Gray (1996), and recently in Mikosch and Stărică (2004), observed high persistence in the conditional

Theis Lange
Department of Economics, University of Copenhagen, Studiestraede 6, 1455 Copenhagen K, Denmark, e-mail: theis.lange@econ.ku.dk

Anders Rahbek
Department of Economics, University of Copenhagen, Studiestraede 6, 1455 Copenhagen K, Denmark, e-mail: anders.rahbek@econ.ku.dk

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_38,

variance implied by single regime models for stock returns and interest rates, might be spuriously caused by parameters varying through the sample, corresponding to different, high and low say, volatility periods. Regime switching models specifically allow for describing this kind of phenomenon.

Numerous regime switching models have been proposed. It is not possible to treat all in detail. The aim of this survey is therefore to convey some of the main ideas. The focus will be on econometric methodology for a few selected models exhibiting regime changes in the conditional mean or variance and their applications. Each regime switching model has its own characteristics. Therefore, and to facilitate the presentation, we first propose a general distinction between 'Markov' and 'observation' switching models. Building upon this, in Section 2 we define switching autoregressive conditional heteroscedastic (ARCH), and cointegrated vector autoregressive (CVAR) models with switching long-run adjustments, and we present some of the dynamic properties. The non-switching versions of these models are central workhorses in time series analysis. Therefore the corresponding switching versions have also received much current interest. Likelihood based estimation and asymptotic theory are commented on in Section 3, while Section 4 briefly addresses hypothesis testing.

1.1 Markov and observation switching

In what follows, X_t is the observed process, returns or exchange rates say, while the finite valued switching process s_t may be observed or unobserved, depending on the model considered for the joint process X_t and s_t. In general, a regime switching model is specified by *(i)* the evolution of X_t, $X_t \in \mathbb{R}^p$, given s_t and $(X_n, s_n)_{n<t}$, and *(ii)* the switching variable s_t given $(X_n, s_n)_{n<t}$.

A useful reference point for *(i)* are models, where the conditional distribution of X_t given the regime s_t and $(X_n, s_n)_{n<t}$ depends only on s_t and the lagged vector $\mathbb{X}_{t-1} = \left(X'_{t-1}, ..., X'_{t-k}\right)'$ for some $k \geq 1$, that is

$$X_t = f_\theta\left(\mathbb{X}_{t-1}, s_t, \epsilon_t\right) \quad \text{for } t = 1, 2, ..., T, \tag{1}$$

where $f_\theta(\cdot)$ are functions indexed by a parameter $\theta \in \Theta$ and the innovations ϵ_t are i.i.d. with a known distribution.

To fix ideas we initially consider two key examples. We commence with regime changing in a univariate ($p = 1$) autoregressive (AR) model of order $k = 1$ given by

$$f_\theta\left(X_{t-1}, s_t, \epsilon_t\right) = \rho_{s_t} X_{t-1} + \sigma\epsilon_t, \tag{2}$$

where ϵ_t is an i.i.d. mean zero and unit variance, denoted i.i.d.(0,1), sequence, and the parameters ρ_i, $i \in \mathbb{S}$, and $\sigma > 0$ are scalars. Equation (2) implies that X_t has autoregressive parameter ρ_i if $s_t = i$. By setting for example

$\rho_1 = 1$ and $|\rho_i| < 1$ for the remaining regimes, this allows in particular for X_t to change between random walk and mean-reversion type behavior.

Second, the autoregressive conditional heteroscedastic (ARCH) model of order one with regime changing is given by

$$f_\theta\left(X_{t-1}, s_t, \epsilon_t\right) = h_t^{1/2}\epsilon_t, \quad h_t = \omega_{s_t} + \alpha_{s_t} X_{t-1}^2, \tag{3}$$

with scalar parameters $\alpha_i \geq 0$ and $\omega_i > 0$. In this case X_t given $s_t = i$ and X_{t-1} has conditional variance $\omega_i + \alpha_i X_{t-1}^2$, and therefore this model allows for X_t to change between high and low volatility regimes as measured by the values of the ARCH parameters.

Next we turn to *(ii)*, the specification of the switching variable s_t. The conditional distribution of s_t is characterized in terms of a parameter λ, $\lambda \in \Lambda$, where in many of the models θ and λ vary in a product set implying that estimation simplifies, see later. It is useful to distinguish between two kinds of switching. One is the Markov switching (MS) type, as introduced in Hamilton (1989, 1990), where s_t is an unobserved stationary ergodic Markov chain on $\mathbb{S}$ with transition probabilities,

$$p_{ij} = P\left(s_t = j \mid s_{t-1} = i\right), \quad i, j \in \mathbb{S}. \tag{4}$$

In this case, λ is chosen as the transition matrix $(p_{ij})_{i,j\in\mathbb{S}}$, where by definition $\sum_{j\in\mathbb{S}} p_{ij} = 1$ for each $i \in \mathbb{S}$.

It will be assumed that ϵ_t in (1) and s_t are conditionally independent, that is

$$\begin{aligned} &P\left(s_t = j, \epsilon_t \in A \mid s_{t-1} = i, s_{t-2}, ..., s_0, X_{t-1}, ...X_1, \mathbb{X}_0\right) \\ &\quad = P\left(s_t = j \mid s_{t-1} = i\right) P\left(\epsilon_t \in A\right) \\ &\quad = p_{ij} P\left(\epsilon_t \in A\right) \end{aligned} \tag{5}$$

for Borel sets $A \subseteq \mathbb{R}^p$ and $j \in \mathbb{S}$. Thus in MS models, the switching variable s_t is exogenous in the sense that there is no feedback from the observed process (X_t) to the switching variable.

In line with the concept of observation driven time series introduced in Cox (1981), the second class considered here will be referred to as observation switching (OS) models. In contrast to MS models, the switching probabilities in OS models are allowed to depend on lagged observed variables, while not on the lagged switching variable. Thus in OS models the switching variable s_t is endogenous in the sense that there is indeed feedback from the observed process (X_t) to the switching variable. The joint dynamics of s_t and ϵ_t for OS models are characterized by,

$$\begin{aligned} &P\left(s_t = j, \epsilon_t \in A \mid s_{t-1}, ..., s_0, X_{t-1}, ...X_1, \mathbb{X}_0\right) \\ &= P\left(s_t = j \mid \mathbb{X}_{t-1}\right) P\left(\epsilon_t \in A\right), \end{aligned} \tag{6}$$

for Borel sets $A \subseteq \mathbb{R}^p$ and $j \in \mathbb{S}$. Here the distribution of s_t conditional on $\mathbb{X}_{t-1}$ is given by the probabilities,

$$q_{tj} \equiv P\left(s_t = j \mid \mathbb{X}_{t-1}\right), \tag{7}$$

for $j \in \mathbb{S}$, with $\sum_{j\in\mathbb{S}} q_{tj} = 1$ and q_{tj} are parameterized in terms of the parameter λ. An example of OS with $r = 2$ regimes (see also Wong and Li (2001), Lanne and Saikkonen (2003), Bec et al. (2005), Bec and Rahbek (2004)), is the one with a logistic specification where

$$\log\left(\frac{q_{t1}}{1-q_{t1}}\right) = \gamma_1 + \gamma_2\|\mathbb{X}_{t-1}\|, \tag{8}$$

with $\lambda = (\gamma_1, \gamma_2) \in \mathbb{R}^2$ and $q_{t2} = 1 - q_{t1}$. In this case, the probability that $s_t = 1$ is increasing in the norm $\|\mathbb{X}_{t-1}\|$ provided $\gamma_2 > 0$, and hence large values of $\|\mathbb{X}_{t-1}\|$ imply a large probability for being in regime 1. Discontinuous self-exciting threshold (TR) models of Tong (1990) are also included, where $\lambda = \gamma$ and,

$$q_{t1} = 1 - q_{t2} = 1\left(\|\mathbb{X}_{t-1}\| \geq \gamma\right). \tag{9}$$

Then for large $\|\mathbb{X}_{t-1}\|$, $s_t = 1$ with probability one. See also Balke and Fomby (1997) for other versions of TR models.

An important difference between MS and OS is in terms of interpretation. In OS models X_t feeds back into the regime. Thus, in contrast to MS models, OS models may provide an interpretation of the switching mechanism in terms of lagged observed variables. Which is deemed most adequate depends on the economic theory or question underlying the econometric analysis of the variables X_t.

Note that it has also been suggested to consider regime switching models, where the transition probabilities may depend on both $\mathbb{X}_{t-1}$ and s_{t-1}, but these are not discussed here, see e.g. Diebold et al. (1994), Filardo and Gordon (1998), Filardo (1994). Note furthermore that OS models may be viewed as generalized mixture models interpreting q_{tj} as mixture probabilities depending on lagged endogenous variables, see e.g. Wong and Li (2000), Bec et al. (2005). Hence OS models relate also to classical pure mixture models where q_{tj} in (7) does not depend on $\mathbb{X}_{t-1}$.

2 Switching ARCH and CVAR

In this section we first present the class of switching ARCH models, together with a brief discussion of switching GARCH models. Secondly, cointegrated vector autoregressive (CVAR) models with switching adjustments are considered. The focus is on formulation of the models, and of their stability properties in terms of geometric ergodicity, stationarity and existence

of moments which have been addressed in the literature. Stability properties provide insight both to the expected dynamic behavior of the processes, as well as necessary background for the asymptotic likelihood based inference addressed later.

2.1 Switching ARCH and GARCH

One of the classical conditional variance models in financial time series analysis is the GARCH model, cf. Bollerslev (1986). Fitted GARCH models often lead to the conclusion of high persistence GARCH, or integrated GARCH (IGARCH), implying that the process is not covariance stationary and multiperiod forecasts of volatility will trend upwards. Also, by definition, the GARCH model has symmetric responses to positive and negative shocks. To address such issues, MS ARCH and GARCH type models have been suggested, see e.g. Cai (1994), Dueker (1997), Gray (1996), Hamilton and Lin (1996), Hamilton and Susmel (1994), Susmel (2000), Francq et al. (2001) and, recently, Haas et al. (2004), Li and Lin (2004), Klaassen (2001). OS ARCH models have been proposed inter alia in Fornari and Mele (1997), Glosten et al. (1993), Li and Lam (1995), Zakoïan (1994), Lanne and Saikkonen (2003), see also Franses and van Dijk (2000) for numerous references.

Note that for simplicity of presentation, the conditional mean part of the observed X_t process is set to zero. As often seen in applications, switching ARCH models are extended to include a constant or autoregressive conditional mean μ_t, by replacing X_t by $X_t - \mu_t$.

2.1.1 Models

Extending the ARCH of order one in (3), a general formulation of univariate ($p = 1$) regime changing ARCH(k) models is given by,

$$X_t = f_\theta\left(\mathbb{X}_{t-1}, s_t, \epsilon_t\right) = h_t^{1/2}\epsilon_t \quad \text{for} \quad t = 1, 2, ..., T, \tag{10}$$

$$h_t = \omega_{s_t} + \sum_{i=1}^{k} \alpha_{is_t} X_{t-i}^2\,, \tag{11}$$

where ϵ_t is i.i.d.(0,1) distributed according to some law, typically Gaussian or $t-$distributed. The ARCH parameters are $\theta = (\omega_1, ..., \omega_k, \alpha_{11}, ..., \alpha_{kr})' \in \Theta = \{\omega_i > 0,\ \alpha_{ij} \geq 0, i = 1, ..., k, j = 1, ..., r\}$. The initial values $\mathbb{X}_0 = (X_0, X_{-1}, ..., X_{-k+1})'$ are assumed to be fixed in the statistical analysis for OS ARCH models, and in addition s_0 is fixed for MS ARCH models. Alternatively, s_0 can be treated as an additional parameter, or since s_t in MS

models is assumed to be a stationary ergodic Markov chain, s_0 can be given the stationary initial distribution.

By definition, the switching mechanism in MS ARCH models is given by the Markov transition probabilities p_{ij} in (4). TR ARCH models use q_{tj} in (7), often formulated in terms of the sign of lagged X_t. With $r = 2$ states this can be exemplified by,

$$\alpha_{is_t} = 1(s_t = 1)\,\alpha_{i1} + 1(s_t = 2)\,\alpha_{i2}, \quad q_{t1} = 1(X_{t-1} > 0) = 1 - q_{t2}, \tag{12}$$

see e.g. Glosten et al. (1993). Thus, negative and positive X_{t-1}, measuring (shocks to) returns, have a different impact on the volatility. For a more general formulation let the sets, or regions, A_j, $j \in \mathbb{S}$, be a disjoint partition of $\mathbb{R}^k$ (k matching the dimension of $\mathbb{X}_{t-1}$) where the partition is parameterized by λ. Then TR specifications as in Gourieroux and Monfort (1992), Cline and Pu (2004) are captured by the general formulation,

$$q_{tj} = 1(\mathbb{X}_{t-1} \in A_j)\,. \tag{13}$$

For example, in (12), $A_1 = \{\mathbb{X} = (X_1, ..., X_k) \in \mathbb{R}^k \mid X_1 > 0\} = A_2^c$, and $\lambda = 0$ is known. Replacing the indicator function in (13) by logistic functions or cumulative distribution functions leads to the generalized mixture ARCH class of models Lanne and Saikkonen (2003), see also Alexander and Lazar (2006), Wong and Li (2001), Hass et al. (2004) for mixture ARCH. A simple example is given if one replaces q_{t1} in (12) by logistic functions as in (8).

Following Gray (1996), Dueker (1997), Klaassen (2001), the most straightforward way to generalize (10) to GARCH, is by defining,

$$h_t = \omega_{s_t} + \sum_{i=1}^{k} (\alpha_{is_t} X_{t-i}^2 + \beta_{is_t} h_{t-i}). \tag{14}$$

However, as pointed out by the quoted authors this formulation leads to a likelihood function with an exponentially (in T) growing number of terms, due to path dependence. To circumvent the problem of an intractable likelihood function Haas et al. (2004) and Lanne and Saikkonen (2003) suggest a specification of the form,

$$h_t = h_{s_t,t}, \quad h_{s_t,t} = \omega_{s_t} + \sum_{i=1}^{k} (\alpha_{is_t} X_{t-i}^2 + \beta_{is_t} h_{s_t,t-i}).$$

such that for each state $j \in \mathbb{S}$, $h_{j,t}$ evolves as a GARCH process. This way the likelihood function is indeed tractable. And conditioning on the initial values $(h_{1,0}, ..., h_{r,0})$, or treating these as parameters to be estimated, this formulation is captured by the formulation in (1).

2.1.2 Properties of switching ARCH processes

For MS ARCH processes conditions for stationarity, existence of moments, autocorrelation functions, and ergodicity have been studied in Francq et al. (2001), Francq and Zakoïan (2005) under the assumption of independence of s_t and the i.i.d.(0,1) ϵ_t sequences. For example in the GARCH(1,1) case of (14) with $r = 2$ regimes, a sufficient condition for stationarity and finite second order moments of X_t can be stated in terms of the spectral radius $\rho(\cdot)$, of a certain matrix:

$$\rho\begin{pmatrix} p_{11}(\alpha_{11}+\beta_{11}) & p_{21}(\alpha_{11}+\beta_{11}) \\ p_{12}(\alpha_{12}+\beta_{12}) & p_{22}(\alpha_{12}+\beta_{12}) \end{pmatrix} < 1. \tag{15}$$

This generalizes the well-known condition for classical GARCH, $\alpha_{11}+\beta_{11} < 1$, and allows for one of the regimes to violate this condition, such that $\alpha_{12} + \beta_{12} \geq 1$ say. Thus, switching between persistent IGARCH, even explosive, and non-persistent volatility regimes is not excluded. The results for OS ARCH type processes are very much dependent on the exact type of specification of s_t. Note however that by definition of OS processes, and unlike MS, $\mathbb{X}_t = (X_t, ..., X_{t-k+1})'$ is a Markov chain. Moreover, the so-called drift criterion (see e.g. Tjøstheim (1990)) can in general be used to establish conditions for geometric ergodicity, and hence for stationarity as well as existence of finite moments of X_t. This is used in e.g. Liu et al. (1997), Carrasco and Chen (2002), Cline and Pu (2004) for OS (G)ARCH type processes. Specifically, from Cline and Pu (2004), it follows that for the TR ARCH process in (12), geometric ergodicity and second order moments are implied by,

$$\tfrac{1}{2}(\alpha_{11}^2 + \cdots + \alpha_{k1}^2 + \alpha_{12}^2 + \cdots + \alpha_{k2}^2) < 1,$$

again allowing for switching between persistent and non-persistent volatility regimes. Finally, similar conditions for second-order stationarity of pure mixture (G)ARCH processes can be found in Haas et al. (2004), Wong and Li (2001).

2.2 Switching CVAR

In the much applied cointegrated VAR (CVAR) models (see e.g. Johansen (2008)) the p-dimensional series X_t adjusts linearly to disequilibria as measured by cointegrated relations. The cointegrated relations are stationary linear combinations of the non-stationary, or integrated, series X_t. As mentioned in the introduction, term-structure and exchange rate data exhibit epochs of both seemingly non-stationary and mean-adjusting behavior, which in terms of the CVAR models correspond to periods with, and periods without, ad-

justments to disequilibria, which therefore cannot be modelled by classic CVAR models. OS cointegrated models in which the adjustment coefficients may switch depending on the cointegrating relations have been applied in Aslanidis and Kouretas (2005), Balke and Fomby (1997), Baum and Karasulu (1998), Bec and Rahbek (2004), Clements and Galvão (2004), Gouveia and Rodrigues (2004), Hansen and Seo (2002), Lo and Zivot (2001), Martens et al. (1998), Tsay (1998) among others, see Dufrenot and Mignon (2002) for more references. Such models are often motivated by transaction costs, or policy intervention arguments. MS CVAR models which extend the MS AR models of Hamilton (1989), Hamilton (1994) are motivated by stochastically changing economic regimes as in Akram and Nymoen (2006), Krolzig et al. (2002), Chow (1998), see also Guidolin and Timmermann (2005) for an MS VAR portfolio application.

2.2.1 Models

We consider now switching adjustments coefficients in the CVAR model with one cointegrated relation. The model is given by,

$$\Delta X_t = \alpha_{s_t} \beta' X_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \, \Delta X_{t-i} + \Omega^{1/2} \epsilon_t, \tag{16}$$

where the innovations ϵ_t are p-dimensional i.i.d. standard normals. In (16) the increments $\Delta X_t = X_t - X_{t-1}$ adjust to the cointegrated relation $\beta' X_{t-1}$ through the p-dimensional adjustment coefficients α_j, $j \in \mathbb{S}$. The parameters are β which is a p-dimensional vector and Γ_i and Ω which are $(p \times p)$-dimensional matrices with Ω positive definite.

Based on Balke and Fomby (1997) and the economic theory of Dumas (1992), Bec and Rahbek (2004) analyzes term structure data using models such as (16), with q_{tj} in (7) a function of the magnitude of $\beta' X_{t-1}$, measuring the spread between interest rates. In case of $r = 2$ regimes this is exemplified by replacing $\mathbb{X}_{t-1}$ by $\beta' X_{t-1}$ in the logistic and TR specifications in (8) and (9). Hence adjustment through α_1 is likely to take place if $||\beta' X_{t-1}||$ is large. The adjustments occurring through α_2 reflect the role of transaction costs and as a limiting case, $\alpha_2 = 0$, and no adjustment takes place if say, the spread $\beta' X_{t-1}$ is negligible. In contrast to the latter model, Hansen and Seo (2002) consider a two regime TR specification, and the q_{tj} in (9) replaced by

$$q_{t1} = 1 \left(\beta' X_{t-1} > \gamma\right) = 1 - q_{t2}. \tag{17}$$

Note that since β enters both (16) and the mentioned OS specifications, it is in general only if β is considered known that the parameters θ and λ vary in a product set.

In Akram and Nymoen (2006), Giese (2006) MS CVAR as in (16) is considered, while in line with Hamilton (1989), Krolzig (1997), the MS model applied in Krolzig et al. (2002) uses level switching rather than switching in the adjustment coefficients.

2.2.2 Properties of switching CVAR processes

It is fundamental for the interpretation of equation (16) that $\beta' X_t$ and ΔX_t are processes which are 'stable', while X_t is not. In the linear case of no switching this distinction is clear-cut as X_t can be shown to be integrated of order one, I(1), while $\beta' X_t$ and ΔX_t are I(0) processes which have stationary representations. In Bec and Rahbek (2004), Saikkonen (2005), Saikkonen (2008) it is used that $\mathbb{X}_t$ is a Markov chain in OS cointegrated VAR models to explore geometric ergodicity, and the implied stationarity, as replacing I(0). Specifically, with $r = 2$ regimes and with q_{t1} increasing to 1 as $||\beta' X_t||$ increases, it is found that if the VAR coefficients of (16) satisfy the well-known cointegration restrictions for regime one only, then $\beta' X_t$ and ΔX_t are stationary. Moreover, they show that X_t appropriately normalized satisfies a functional central limit theorem, and is in this sense non-stationary. What is important here is that there are no restrictions on α_2, such that indeed α_2 may be zero and hence accommodating for transaction costs as desired. For more general versions, including the TR case as in (17), Saikkonen (2008) finds that the general concept of joint spectral radius involving coefficients for all regimes is needed to state sufficient conditions for geometric ergodicity.

In MS CVAR models, $\mathbb{X}_t$ is not a Markov chain while $(\mathbb{X}_t, s_t)$ is. This is used for studies of geometric ergodicity in general MS VAR models in Lee (2005), Yao (2000) which may be used for the CVAR case, see also Ulloa (2006). Results for stationarity, ergodicity and existence of moments for MS (V)AR models can be found in Francq and Zakoïan (2001), Francq and Roussignol (1998), Yang (2000), Yao (2001).

3 Likelihood–Based Estimation

Next consider likelihood-based estimation of the parameters in switching models given by (1) and either (6) or (5), that is OS or MS specifications. Denote by $g_\theta(X_t|s_t = j, \mathbb{X}_{t-1})$ the conditional density of X_t given $s_t = j$ and $\mathbb{X}_{t-1}$, indexed by the parameter vector θ.

Turn first to OS models, where $q_{tj} = P(s_t = j|\mathbb{X}_{t-1})$, $j \in \mathbb{S}$, with q_{tj} parametrized by λ. Then, conditionally on $\mathbb{X}_0$, the likelihood function for the OS models can be written as,

$$L(X_1, ..., X_T; \theta, \lambda | \mathbb{X}_0) = \prod_{t=1}^{T} \left\{ \sum_{j=1}^{r} g_\theta(X_t | s_t = j, \mathbb{X}_{t-1})\, q_{tj} \right\}. \tag{18}$$

In TR models, the regime process s_t is observable, the q_{tj} are indicator functions, and with θ and λ varying freely, maximization of the log-likelihood function in (18) is typically solved by letting λ vary over a grid, and maximizing over θ for each value of λ. See Hansen (2000), Tong (1990) for autoregressions and Liu et al. (1997), Tsay (1998), where ARCH is included. See also Hansen and Seo (2002) for the TR CVAR case, where λ and θ do not vary freely.

For general OS models one may iteratively update the likelihood function by the EM algorithm (see Ruud (1997) for an introduction) and thereby obtain the estimators $\widehat{\theta}$ and $\widehat{\lambda}$. To maximize the log-likelihood function in (18) introduce $\mathcal{X} = (X_1, ..., X_T)$ and $\mathcal{S} = (s_1, ..., s_T)$, and consider the full log-likelihood function for all observations $\mathcal{X}$ and $\mathcal{S}$, treating the latter as observed, and with $\mathbb{X}_0$ fixed,

$$\log L(\mathcal{X}, \mathcal{S} | \mathbb{X}_0) = \sum_{t=1}^{T} \sum_{j=1}^{r} \{ \log g_\theta(X_t | s_t = j, \mathbb{X}_{t-1}) + \log q_{tj} \} 1(s_t = j).$$

Then, corresponding to the E-step of the algorithm,

$$\begin{aligned} &E[\log L(\mathcal{X}, \mathcal{S} | \mathbb{X}_0) | \mathcal{X}] \\ &= \sum_{t=1}^{T} \sum_{j=1}^{r} \pi_{jt} \log g_\theta(X_t | s_t = j, \mathbb{X}_{t-1}) + \sum_{t=1}^{T} \sum_{j=1}^{r} \pi_{jt} \log q_{tj}, \end{aligned} \tag{19}$$

where $\pi_{jt} = E(1(s_t = j) | \mathcal{X})$ are the 'smoothed probabilities' given by,

$$\begin{aligned} \pi_{jt} &= P(s_t = j | \mathcal{X}) = P(s_t = j | X_t, \mathbb{X}_{t-1}) \\ &= \frac{q_{tj} g_\theta(X_t | \mathbb{X}_{t-1}, s_t = j)}{\sum_{i=1}^{r} q_{ti} g_\theta(X_t | \mathbb{X}_{t-1}, s_t = i)}. \end{aligned}$$

The nth step of the iterative optimization is given by fixing π_{jt} at $(\theta, \lambda) = (\widehat{\theta}^{(n-1)}, \widehat{\lambda}^{(n-1)})$, and then, if θ and λ vary freely, maximizing each term in (19) separately over θ and λ to obtain $\widehat{\theta}^{(n)}$ and $\widehat{\lambda}^{(n)}$, respectively. See Bec et al. (2005), Wong and Li (2000) and Wong and Li (2001) for application of the EM algorithm to (generalized) mixture autoregressions and Bec and Rahbek (2004) for CVAR.

For MS models it is straightforward to write down the likelihood function in terms of $g_\theta(X_t | s_t = j, \mathbb{X}_{t-1})$ and the switching probabilities $p_{ij} = P(s_t = j | s_{t-1} = i)$ similar to (18), see also Yang (2001). However, due to computational length and complexity, predominantly iterative algorithms such as the EM algorithm are used to optimize the likelihood function. Similar to the

OS case, the full likelihood, fixing $\mathbb{X}_0$ and s_0, is given by,

$$\begin{aligned}&\log L\left(\mathcal{X},\mathcal{S}|\mathbb{X}_0,s_0\right)\\&=\sum_{t=1}^{T}\sum_{i,j=1}^{r}\left\{\log g_\theta\left(X_t|s_t=j,\mathbb{X}_{t-1}\right)+\log p_{ij}\right\}1\left(s_t=j,s_{t-1}=i\right).\end{aligned}$$

The expectation of the above given $\mathcal{X}$ is identical to (19), however with the crucial difference that the smoothed probabilities are replaced by

$$\pi_{ij,t}=P\left(s_t=j,s_{t-1}=i|\mathcal{X}\right).$$

No simple closed form solution exist for $\pi_{ij,t}$. Instead these have to be computed for example by the so-called forward and backward recursions, see Hamilton (1990), Holst et al. (1994), and more generally McLachlan and Peel (2000), for details. See also Hamilton (1989), Hamilton (1994), Hamilton and Raj (2002), Kim (2004), Krolzig (1997) for further discussions on optimization algorithms in MS models.

As to asymptotic theory for maximum likelihood estimators the literature is not complete yet, but some results can be emphasized. For TR models, in general the discontinuities implied by the threshold lead to non-standard limiting distributions of $\widehat{\lambda}$, see Chan (1993), Hansen (1997), Hansen (2000). On the other hand, $\widehat{\theta}$ is in general asymptotically Gaussian distributed, see Liu et al. (1997), Tsay (1998) for TR autoregressive and ARCH models. See also Gourieroux and Monfort (1992), Kristensen and Rahbek (2005) for results in some special cases of TR ARCH models. For OS CVAR models the asymptotic distributions of the cointegration vector β and error-correction parameters is given in Kristensen and Rahbek (2007), see also Bec and Rahbek (2004) for known β. Note that some recent results for smoothed non-maximum likelihood estimation of β in a TR CVAR model can be found in Seo (2006), see also Jong (2002).

For MS models Francq et al. (2001) show consistency of MS ARCH parameters, while Krishnamurthy and Rydén (1998) consider MS AR. Under general regularity conditions, a complete asymptotic distribution theory for MS models of the form in (1) has recently been given by Douc et al. (2004), see also Fuh (2004).

4 Hypothesis Testing

Testing for the number r of regimes, or states, which s_t switches between is of main concern in switching models. This includes in particular tests of whether switching is suitable at all, that is $r = 1$. In terms of the emphasized examples this would correspond to the classic ARCH and linear CVAR models rather than their switching counterparts. In general, the hypothesis, H_0, of

interest is formulated as a parametric restriction on θ in (1), which implies that λ (which parametrize the switching probabilities for s_t, cf. (4) and (7) respectively for MS and OS) is unidentified.

As an example, consider the case of the AR(1) model in (2) for X_t, with $r = 2$ regimes such that $\theta = (\rho_1, \rho_2, \sigma)$. The hypothesis of a single regime, or linearity, may then be represented as $H_0 : \rho_1 = \rho_2$, under which $\lambda = (p_{ij})_{i,j=1,2} \in \Lambda$ is unidentified in the case of MS, while $\lambda = \gamma \in \Lambda$ is unidentified in the case of TR specifications as in (9) or (17).

It should be emphasized that the lack of identification of λ under H_0 does not mean that the likelihood ratio test statistic (LR) for the hypothesis H_0 cannot be computed. Specifically, the likelihood function can be maximized as discussed in Section 3 under the alternative, while maximization under H_0 is standard as in the above AR(1) case where estimation reduces to ordinary regression. However, as discussed in inter alia Andrews (1993), Davies (1977), Davies (1987), Hansen (1992), Hansen (1996), standard asymptotic theory often does not apply to the LR when λ is unidentified under the null H_0. In fact, even in the simple classic case of Gaussian mixture models, where the asymptotic distribution of the LR 'has long been a mystery' (p. 62 Liu and Shao (2004)), the same authors find that the LR diverges to infinity at the rate of $\log(\log T)$, see also Corollary 1 in Andrews (1993).

Much interest has therefore been devoted to the widely applied so-called 'sup' class of tests which, under regularity conditions, indeed have limiting distributions for which asymptotic p-values and critical values can be obtained by simulations, see e.g. Hansen (1996) for a discussion of bootstrap related techniques. In terms of the AR(1) example above, denote by $\mathrm{LR}(\lambda)$ the likelihood ratio test for H_0 for a fixed λ. Then the 'sup' version of the LR is given by

$$\sup_{\lambda \in \widetilde{\Lambda}} \mathrm{LR}(\lambda),$$

where $\widetilde{\Lambda}$ is a suitably chosen compact subset of Λ. The limiting distribution of the supLR statistic is given in terms of Gaussian processes in Chan (1990), Chan and Tong (1990) for the TR case, while for the MS case it is discussed in Garcia (1998), see also Carrasco (2002) for a joint discussion. For practical purposes, then as in Hansen (1996), $\widetilde{\Lambda}$ is chosen for the TR case such that e.g. 5% of the observations of X_t lie in each regime, which in particular implies that $P(s_t = 1) = 1 - P(s_t = 2)$ is strictly bounded away from 0 and 1. For the MS case $\widetilde{\Lambda}$ is likewise chosen such that $P(s_t = 1) = p_{21}/(1 - p_{11} + p_{21})$ is strictly bounded away from 0 and 1, see Carrasco (2002), Garcia (1998). Note that 'sup' versions of Lagrange multiplier (LM) and Wald type tests are also widely applied, see Altissimo and Corradi (2002).

In the mentioned references an underlying assumption is that under H_0 the analyzed process is stationary. For the non-stationary case, Hansen and Seo (2002) studies the supLM test for the null of linear cointegration. These results extend testing the null of a univariate integrated non-stationary pro-

cess, see e.g. Bec et al. (2004), Caner and Hansen (2001), Enders and Granger (1998).

For another strand of literature, largely based on information criteria, addressing the selection of the number of regimes, see inter alia Francq et al. (2001), Hass et al. (2004) for switching ARCH models, Psadarakis and Spagnolo (2003), Wong and Li (2001) for MS and mixture AR models, respectively, see also Gonzalo and Pitarakis (2002) for a different approach to determination of regimes in TR VAR models.

5 Conclusion

From the presented results and references on regime switching models, one may conclude that, while the models receive much interest in applications, a sufficiently complete theory for these models is still missing. An issue which has not been addressed in this survey is the use of the models for forecasting. For an introduction, see inter alia Franses and van Dijk (2000) and the references therein, and Amendola and Niglio (2004), Davidson (2004), Clements and Krolzig (1998) for recent discussions of forecasting in TR and MS models.

References

Akram, Q. and Nymoen, R. (2006): Econometric modelling of slack and tight labour markets. *Economic Modelling* **23**, 579–596.

Alexander, C. and Lazar, E. (2006): Normal mixture GARCH(1,1): Applications to exchange rate modelling. *Journal of Applied Econometrics* **21**, 307–336.

Altissimo, F. and Corradi, V. (2002): Bounds for inference with nuisance parameters present only under the alternative. *The Econometrics Journal* **5**, 494–519.

Amendola, A. and Niglio, M. (2004): Predictor distribution and forecast accuracy of threshold models. *Statistical Methods & Applications* **13**, 3–14.

Andrews, D. (1993): Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**, 821–856.

Aslanidis, N. and Kouretas, G. (2005): Testing for two-regime threshold cointegration in the parallel and official markets for foreign currency in Greece. *Economic Modelling* **22**, 665–682.

Balke, N. and Fomby, T. (1997): Threshold cointegration. *International Economic Review* **38**, 627–645.

Baum, C. and Karasulu, M. (1998): Modelling Federal Reserve discount policy. *Computational Economics* **11**, 53–70.

Bec, F. and Rahbek, A. (2004): Vector equilibrium correction models with non-linear discontinuous adjustments. *The Econometrics Journal* **7**, 628–651.

Bec, F., Rahbek, A. and Shephard, N. (2005): The ACR Model: A Multivariate Dynamic Mixture Autoregression. *Oxford Bulletin of Economics and Statistics* forthcoming.

Bec, F., Salem, M.B. and Carrasco, M. (2004): Tests for unit-root versus threshold specification with an application to the purchasing power parity relationship. *Journal of Business and Economic Statistics* **22**, 382–395.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Cai, J. (1994): A Markov model of switching-regime ARCH. *Journal of Business and Economic Statistics* **12**, 309–316.

Caner, M. and Hansen, B. (2001): Threshold autoregression with a unit root. *Econometrica* **69**, 1555–1596.

Carrasco, M. (2002): Misspecified structural change, threshold, and Markov-switching models. *Journal of Econometrics* **109**, 239–273.

Carrasco, M. and Chen, X. (2002): Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17–39.

Chan, K. (1990): Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society. Series B* **53**, 691–696.

Chan, K. (1993): Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of Statistics* **21**, 520–533.

Chan, K. and Tong, H. (1990): On likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society. Series B* **53**, 469–476.

Chow, Y. (1998): Regime switching and cointegration tests of the efficiency of futures markets. *Journal of Futures Markets* **18**, 871–901.

Clements, M. and Galvão, A. (2004): Testing the expectations theory of the term structure of interest rates in threshold models. *Macroeconomic Dynamics* **7**, 567–585.

Clements, M. and Krolzig, H. (1998): A comparison of the forecast performance of markov-switching and threshold autoregressive models of US GNP. *The Econometrics Journal* **1**, C47–C75.

Cline, D. and Pu, H. (2004): Stability and the Lyapounov exponent of threshold AR-ARCH models. *Annals of Applied Probability* **14**, 1920–1949.

Cox, D. (1981): Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.

Davidson, J. (2004): Forecasting Markov-switching dynamic, conditionally heteroscedastic processes. *Statistics and Probability Letters* **68**, 137–147.

Davies, R. (1977): Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.

Davies, R. (1987): Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

Diebold, F., Lee, J. and Weinbach, G. (1994): Nonstationary Time Series Analysis and Cointegration. *Advanced Texts in Econometrics*, 283–302. Oxford University Press.

Douc, R., Moulines, E. and Rydén, T. (2004): Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics* **32**, 2254–2304.

Dueker, M. (1997): Markov switching in GARCH processes and mean-reverting stock-market volatility. *Journal of Business and Economic Statistics* **15**, 26–34.

Dufrenot, G. and Mignon, V. (2002): *Recent Developments in Nonlinear Cointegration with Applications to Macroeconomics and Finance*. Kluwer Academic Publishers

Dumas, B. (1992): Dynamic equilibrium and the real exchange rate in a spatially seperated world. *Review of Financial Studies* **5**, 153–180.

Enders, W. and Granger, C. (1998): Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates. *Journal of Business and Economic Statistics* **16**, 304–311.

Filardo, A. (1994): Business-cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics* **12**, 299–308.

Filardo, A. and Gordon, S. (1998): Business cycle durations. *Journal of Econometrics* **85**, 99–123.

Fornari, F. and Mele, A. (1997): Sign- and volatility-switching ARCH models: Theory and applications to international stock markets. *Journal of Applied Econometrics* **12**, 49–65.

Francq, C. and Roussignol, M. (1998): Ergodicity of autoregressive processes with Markov switching and consistency of the maximum-likelihood estimator. *Statistics* **32**, 151–173.

Francq, C., Roussignol, M. and Zakoïan, J. (2001): Conditional heteroskedasticity driven by hidden Markov chains. *Journal of Time Series Analysis* **22**, 197–220.

Francq, C. and Zakoïan, J. (2001): Stationarity of multivariate Markov-switching ARMA models. *Journal of Econometrics* **102**, 339–364.

Francq, C. and Zakoïan, J. (2005): The L^2-structures of standard and switching-regime GARCH models. *Stochastic Processes and their Applications* **115**, 1557–1582.

Franses, P. and van Dijk, D. (2000): *Nonlinear Time Series Models in Empirical Finance*. Cambridge University Press.

Fuh, C. (2004): On Bahadur efficiency of the maximum likelihood estimator in hidden Markov models. *Statistica Sinica* **14**, 127–155.

Garcia, R. (1998): Asymptotic null distribution of the likelihood ratio test in Markov switching models. *International Economic Review* **39**, 763–788.

Giese, J. (2006): Characterising the yield curve's derivatives in a regime-changing cointegrated VAR model. *Working paper, Nuffield College University of Oxford*.

Glosten, L., Jaganathan, R. and Runkle, D. (1993): On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* **48**, 1779–1801.

Gonzalo, J. and Pitarakis, J. (2002): Estimation and model selection based inference in single and multiple threshold models. *Journal of Econometrics* **110**, 319–352.

Gourieroux, C. and Monfort, A. (1992): Qualitative threshold ARCH models. *Journal of Econometrics* **52**, 159–199.

Gouveia, P. and Rodrigues, P. (2004): Threshold cointegration and the PPP hypothesis. *Journal of Applied Statistics* **31**, 115–127.

Gray, S. (1996): Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* **42**, 27–62.

Guidolin, M. and Timmermann, A. (2005): Economic implications of bull and bear regimes in UK stock and bond returns. *The Economic Journal* **115**, 111–143.

Haas, M., Mittnik, S. and Paolella, M. (2004): Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics* **2**, 211–250.

Haas, M., Mittnik, S. and Paolella, M. (2004): A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics* **2**, 493–530.

Hamilton, J. (1989): A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384.

Hamilton, J. (1990): Analysis of time series subject to changes in regime. *Journal of Econometrics* **45**, 39–70.

Hamilton, J. (1994): *Time Series Analysis*. Princeton University Press.

Hamilton, J. and Lin, G. (1996): Stock market volatility and the business cycle. *Journal of Applied Econometrics* **11**, 573–593.

Hamilton, J. and Raj, B. (2002): *Advances in Markov-switching Models: Applications in Business Cycle Research and Finance*. Physica-Verlag.

Hamilton, J. and Susmel, R. (1994): Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics* **64**, 307–333.

Hansen, B. (1992): The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics* **7**, S61–S82.

Hansen, B. (1996): Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64**, 413–430.

Hansen, B. (1997): Inference in TAR models. *Studies in Nonlinear Dynamics and Econometrics* **2**, 1–14.

Hansen, B. (2000): Sample splitting and threshold estimation. *Econometrica* **68**, 575–603.

Hansen, B. and Seo, B. (2002): Testing for two-regime threshold cointegration in vector error-correction models. *Journal of Econometrics* **110**, 293–318.

Holst, U., Lindgren, G., Holst, J. and Thuvesholmen, M. (1994): Recursive estimation in switching autoregressions with a Markov regime. *Journal of Time Series Analysis* **15**, 489–506.

Johansen, S. (2008): Cointegration: Overview and development. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 671–693. Springer, New York.

Jong, R. de (2002): Nonlinear minimization estimators in the presence of cointegrating relations. *Journal of Econometrics* **110**, 241–259.

Kim, C. (2004): Markov-switching models with endogenous explanatory variables. *Journal of Econometrics* **122**, 127–136.

Klaassen, F. (2001): Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Economics* **27**, 363–394.

Krishnamurthy, V. and Rydén, T. (1998): Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of Time Series Analysis* **19**, 291–307.

Kristensen, D. and Rahbek, A. (2005): Asymptotics of the QMLE for a class of ARCH(q) models. *Econometric Theory* **21**, 946–961.

Kristensen, D. and Rahbek, A. (2007): Likelihood-based inference in Non-linear Error-Correction Models. *Working paper, University of Copenhagen and CREATES*.

Krolzig, H. (1997): *Markov-Switching Vector Autoregressions*. Springer, Berlin.

Krolzig, H., Marcellino, M. and Mizon, G. (2002): A Markov-switching vector equilibrium correction model of the UK labour market. *Empirical Economics* **27**, 233–254.

Lamoureux, C. and Lastrapes, W. (1990): Persistence in variance, structural change, and the GARCH model. *Journal of Business and Economic Statistics* **8**, 225–234.

Lanne, M. and Saikkonen, P. (2003): Modeling the U.S. short-term interest rate by mixture autoregressive processes. *Journal of Financial Econometrics* **1**, 96–125.

Lee, O. (2005): Probabilistic properties of a nonlinear ARMA process with Markov switching. *Communications in Statistics: Theory and Methods* **34**, 193–204.

Li, M. and Lin, H. (2004): Estimating Value-at-Risk via Markov switching ARCH models - an empirical study on stock index returns. *Applied Economics Letters* **11**, 679–691.

Li, W. and Lam, K. (1995): Modelling asymmetry in stock returns by a threshold autoregressive conditional heteroscedastic model. *Statistician* **44**, 333–341.

Liu, J., Li, W. and Li, C. (1997): On a threshold autoregression with conditional heteroscedastic variances. *Journal of Statistical Planning and Inference* **62**, 279–300.

Liu, X. and Shao, Y. (2004): Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference* **123**, 61–81.

Lo, M. and Zivot, E. (2001): Threshold cointegration and nonlinear adjustment to the law of one price. *Macroeconomic Dynamics* **5**, 506–532.

Martens, M., Kofman, P. and Vorst, T. (1998): A threshold error-correction model for intraday futures and index returns. *Journal of Applied Econometrics* **13**, 245–263.

McLachlan, G. and Peel, D. (2000): *Finite Mixture Models. Wiley Series in Probability and Statistics*. Wiley, New York.

Mikosch, T. and Stărică, C. (2004): Non-stationarities in financial time series, the long range dependence and the IGARCH effects. *Review of Economics and Statistics* **86**, 378–390.

Psadarakis, Z. and Spagnolo, N. (2003): On the determination of the number of regimes in Markov-switching autoregressive models. *Journal of Time Series Analysis* **24**, 237–252.

Ruud, P. (1997): Extensions of estimation methods using the EM algorithm. *Journal of Econometrics* **49**, 305–341.

Saikkonen, P. (2005): Stability results for nonlinear error correction models. *Journal of Econometrics* **127**, 69–81.

Saikkonen, P. (2008): Stability of regime switching error correction models. *Econometric Theory* forthcoming.

Seo, M. (2006): *Estimation of threshold cointegration. London School of Economics, unpublished manuscript.*

Susmel, R. (2000): Switching volatility in private international equity markets. *International Journal of Finance and Economics* **5**, 265–283.

Tjøstheim, D. (1990): Non-linear time series and Markov chains. *Advances in Applied Probability* **22**, 587–611.

Tong, H. (1990): *Non-Linear Time Series. Oxford Statistical Science Series.* Oxford University Press, Oxford.

Tsay, R. (1998): Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* **93**, 1188–1202.

Ulloa, R. (2006): *Essays in Nonlinear Time Series.* Ph.D. thesis, University of Warwick, Department of Economics.

Wong, C. and Li, W. (2000): On a mixture autoregressive model. *Journal of the Royal Statistical Society, Series B* **62**, 95–115.

Wong, C. and Li, W. (2001): On a logistic mixture autoregressive model. *Biometrika* **88**, 833–846.

Wong, C. and Li, W. (2001): On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association* **96**, 982–995.

Yang, M. (2000): Some properties of vector autoregressive processes with Markov-switching coefficients. *Econometric Theory* **16**, 23–43.

Yang, M. (2001): Closed-form likelihood function of Markov-switching models. *Economics Letters* **70**, 319–326.

Yao, J. (2001): On square-integrability of an AR process with Markov switching. *Statistics & Probability Letters* **52**, 265–270.

Yao, J. and Attali, J. (2000): On stability of nonlinear AR processes with Markov switching. *Advances in Applied Probability* **32**, 394–407.

Zakoïan, J. (1994): Threshold heteroskedastic models. *Journal of Economic Dynamics and Control* **18**, 931–955.

Model Selection

Hannes Leeb and Benedikt M. Pötscher

Abstract We provide an overview of the vast and rapidly growing area of model selection in statistics and econometrics.

1 The Model Selection Problem

Model selection has become a ubiquitous statistical activity in the last few decades, none the least owing to the computational ease with which many statistical models can be fitted to data with the help of modern computing equipment. In this article we provide an introduction to the statistical aspects and implications of model selection and we review the relevant literature.

1.1 A general formulation

When modeling data Y, a researcher often has available a menu of competing candidate models which could be used to describe the data. Let $\mathcal{M}$ denote the collection of these candidate models. Each model M, i.e., each element of $\mathcal{M}$, can—from a mathematical point of view—be viewed as a collection of probability distributions for Y implied by the model. That is, M is given by

$$M = \{\mathbb{P}_\eta : \eta \in H\},$$

Hannes Leeb
Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA, e-mail: hannes.leeb@yale.edu

Benedikt M. Pötscher
Department of Statistics, University of Vienna, Universitätsstrasse 5, 1010 Vienna, Austria, e-mail: benedikt.poetscher@univie.ac.at

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_39,

where $\mathbb{P}_\eta$ denotes a probability distribution for Y and H represents the "parameter" space (which can be different across different models M). The parameter space H need not be finite-dimensional. Often, the "parameter" η will be partitioned into (η_1, η_2), where η_1 is a finite-dimensional parameter, whereas η_2 is infinite-dimensional. In the case where the parameterization is identified, i.e., the map $\eta \to \mathbb{P}_\eta$ is injective on H, we will often not distinguish between M and H and will use them synonymously.

The model selection problem is now to select—on the basis of the data Y—a model $\widehat{M} = \widehat{M}(Y)$ in $\mathcal{M}$ such that $\widehat{M}$ is a "good" model for the data Y. Of course, the sense in which the selected model should be a "good" model needs to be made precise and is a crucial point in the analysis. This is particularly important if—as is usually the case—selecting the model $\widehat{M}$ is not the final purpose of the analysis, but $\widehat{M}$ is used as a basis for the construction of parameter estimators, predictors, or other inference procedures.

Typically, with each model M we will have associated an estimator $\widehat{\eta}(M)$ such as the maximum likelihood estimator or the least squares estimator, etc. It is important to note that when model selection precedes parameter estimation, the estimator finally reported is $\widetilde{\eta} = \widehat{\eta}(\widehat{M})$ (and *not* one of the estimators $\widehat{\eta}(M)$). We call $\widetilde{\eta}$ a post-model-selection estimator (PMSE). It is instructive to note that $\widetilde{\eta}$ can be written as

$$\widetilde{\eta} = \sum_{M \in \mathcal{M}} \widehat{\eta}(M) \mathbf{1}(\widehat{M} = M),$$

which clearly shows the compound nature of the PMSE. Note that the above sum is well defined even if the spaces H for different M bear no relationship to each other.

We note that in the framework just described it may or may not be the case that one of the candidate models M in $\mathcal{M}$ is a correct model (in the sense that the actual distribution of the data coincides with a distribution $\mathbb{P}_\eta$ in M). A few examples illustrating the above notation are in order.

Example 1 (selection of regressors)
Suppose Y is an $n \times 1$ vector generated through

$$Y = X\theta + u, \tag{1}$$

where X is an $n \times K$ matrix of nonstochastic regressors with full column-rank and u is a disturbance term whose distribution F does not depend on θ and varies in a set $\mathcal{F}$ of distributions (e.g., $\mathcal{F}$ could be the set of all $N(0, \sigma^2 I_n)$ distributions). Suppose the researcher suspects that some regressors (i.e., columns of X) are superfluous for explaining Y (in the sense that the true values of the coefficients of these regressors are zero), but does not know which of the regressors are superfluous. Then the appropriate candidate models are all submodels of (1) given by zero restrictions on the parameter vector θ. More formally, let $\mathfrak{r} \in \{0,1\}^K$, i.e., $\mathfrak{r}$ is a $K \times 1$ vector of zeros and ones.

Then each $\mathfrak{r} \in \{0,1\}^K$ defines a submodel

$$M_{\mathfrak{r}} = \left\{(\theta, F) \in \mathbb{R}^K \times \mathcal{F} : \theta_i = 0 \text{ if } \mathfrak{r}_i = 0\right\},$$

the full model M_{full} corresponding to $\mathfrak{r} = (1, \ldots, 1)$. The set of all candidate models is given by

$$\mathcal{M}_{all} = \left\{M_{\mathfrak{r}} : \mathfrak{r} \in \{0,1\}^K\right\}.$$

The set-up just described could be termed "all-subset selection." If—on a priori grounds—one wants to protect some of the variables, say, the first k ones, from being eliminated by the model selection procedure, one would then of course consider as candidate models only those in the set

$$\mathcal{M}_{protected} = \left\{M_{\mathfrak{r}} : \mathfrak{r} \in \{0,1\}^K,\ \mathfrak{r}_i = 1 \text{ for } i = 1, \ldots, k\right\}.$$

Another case arises if there is an a priori given ordering of the regressors reflecting their perceived "importance" in explaining Y. For example, in polynomial regression one would usually include a certain power of the explanatory variable only if all lower-order terms are also included. If we assume, without loss of generality, that the ordering of the columns of the matrix X reflects the a priori given ordering, then this amounts to considering

$$\mathcal{M}_{nested} = \{M(p) : 0 \leq p \leq K\}$$

as the set of candidate models, where

$$M(p) = \left\{(\theta, F) \in \mathbb{R}^K \times \mathcal{F} : \theta_i = 0 \text{ for } i > p\right\}.$$

Note that in this case the models $M(p)$ are nested in the sense that $M(p) \subseteq M(p+1)$ holds. Yet another variant is the set of candidate models

$$\mathcal{M}_{nested,protected} = \{M(p) : k \leq p \leq K\}$$

which obviously protects the first k variables in the context of nested model selection. If M is now a submodel of (1), one would typically estimate the parameters of model M by the (restricted) least squares estimator $\widehat{\theta}(M)$ associated with M. Given a model selection procedure $\widehat{M}$ selecting from a set $\mathcal{M}$ of candidate models, the associated PMSE is then given by

$$\widetilde{\theta} = \sum_{M \in \mathcal{M}} \widehat{\theta}(M) \mathbf{1}(\widehat{M} = M). \tag{2}$$

For an extension of this example to the case of infinitely many regressors see Section 3.

Example 2 (linear restrictions)
Suppose the overall model is again given by model (1) but the submodels are now defined by general linear restrictions of the form $R\theta = r$.

Example 3 (time series models)
Suppose the data $Y = (y_1, \ldots, y_n)'$ follow an autoregressive model

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_P y_{t-P} + u_t$$

for $t \geq 1$ and initial values $y_0, \ldots, y_{1-P}$. Typical assumptions on the errors u_t are that they are independent and identically distributed according to a distribution with mean zero, or that the errors form a martingale difference sequence, etc. Of interest here are those submodels where $\theta_{p+1} = \theta_{p+2} = \ldots = \theta_P = 0$, in which case the model selection problem is the problem of selecting the order of the autoregressive process. [Similarly, the order selection problem for other classes of time series models such as, e.g., autoregressive moving average models or generalized autoregressive conditional heteroscedasticity (GARCH) models obviously also fits into the framework outlined above.] In this example we have assumed that y_t is generated by a finite-order autoregressive model. Often finite-order autoregressive models are fitted to a time series, e.g., for the purpose of prediction, even if the time series is not a finite-order autoregression. In this case the order of the approximating autoregressive model has to be determined from the data, leading again to a model selection problem that falls under the umbrella of the general framework formulated above.

Example 4 (general parametric models)
Starting from an overall parametric model $\{\mathbb{P}_\eta : \eta \in H\}$, submodels M_g are defined by restrictions $g(\eta) = 0$, i.e., $M_g = \{\mathbb{P}_\eta : \eta \in H,\ g(\eta) = 0\}$, for g belonging to a given class $\mathcal{G}$ of restrictions. The set $\mathcal{M}$ is then given by $\mathcal{M} = \{M_g : g \in \mathcal{G}\}$. Note that models corresponding to different restrictions g will in general not be nested in each other, although they are nested in the overall model.

1.2 Model selection procedures

1.2.1 Procedures based on tests

Consider for simplicity first the case of only two competing candidate models M_i, $i = 1, 2$, where one is nested in the other, e.g., $M_1 \subseteq M_2$. Furthermore, assume that at least the larger model M_2 is correct, i.e., that the true probability distribution of Y belongs to M_2. Then a decision between the models M_1 and M_2 can be based on a test of the hypothesis H_0 that the true probability distribution belongs to M_1 versus the alternative H_1 that it belongs to $M_2 \backslash M_1$. More formally, let $\mathfrak{R}$ be a rejection region of a test for the hypothesis H_0. Then the selected model $\widehat{M}$ is given by

$$\widehat{M} = \begin{cases} M_1 \text{ if } Y \notin \mathfrak{R} \\ M_2 \text{ if } Y \in \mathfrak{R} \end{cases}.$$

For example, if M_2 corresponds to the linear model (1) with independent identically $N(0, \sigma^2)$ distributed errors and M_1 is given by a linear restriction $R\theta = r$, then the rejection region $\mathfrak{R}$ could be chosen as the rejection region of a classical F test of this linear restriction.

In the case of more than two candidate models which are nested, i.e., $M_1 \subseteq M_2 \subseteq \ldots \subseteq M_s$ holds, model selection can be based on a sequence of tests. For example, one can start by testing M_{s-1} against M_s. If this test results in rejection, one sets $\widehat{M} = M_s$. Otherwise, a test of M_{s-2} against M_{s-1} is performed. If this second test results in rejection, one sets $\widehat{M} = M_{s-1}$. If this second test does not result in rejection, one proceeds with testing M_{s-3} against M_{s-2} and so on, until a test results in rejection or one has reached the smallest model M_1. Such a procedure is often called a "general-to-specific" procedure. Of course, one could also start from the smallest model and conduct a "specific-to-general" testing procedure. If the set $\mathcal{M}$ of candidate models is not ordered by the inclusion relation ("nonnested case"), testing procedures can still be used to select a model $\widehat{M}$ from $\mathcal{M}$, although then more thought has to be given to the order in which to conduct the tests between competing models. The familiar stepwise regression procedures (see, e.g., Chapter 6 in Draper and Smith (1981) or Hocking (1976)) are a case in point. Model selection procedures based on hypothesis tests have been considered, e.g., in Anderson (1962, 1963), McKay (1977), Pötscher (1983, 1985), Bauer et al. (1988), Hosoya (1984, 1986), and Vuong (1989); for a more recent contribution see Bunea et al. (2006). Also the related literature on pretest estimators as summarized in Bancroft and Han (1977), Judge and Bock (1978), and Giles and Giles (1993) fits in here.

Returning to the case of two nested models M_1 and M_2, we note that the model selection procedures sketched above are based on testing whether the true distribution of Y belongs to model M_1 or not. However, if the goal is not so much selection of the "true" model but is selection of a model that results in estimators with small mean squared error, it may be argued that the appropriate hypothesis to test is not the hypothesis that the distribution of Y belongs to M_1, but rather is the hypothesis that the mean squared error of the estimator based on M_1 is smaller than the mean squared error of the estimator based on M_2. Note that this is not the same as the hypothesis that the distribution of Y belongs to M_1. This observation seems to have first been made by Toro-Vizcarrondo and Wallace (1968); see also Wallace (1972). In the context where M_2 is a normal linear regression model and M_1 is given by a linear restriction $R\theta = r$, they showed that the mean squared error matrix of the restricted least squares estimator is less than or equal to the mean squared error matrix of the unrestricted least squares estimator whenever $\sigma^{-2}\theta' R'[R(X'X)^{-1}R']^{-1}R\theta \leq 1$ holds. Hence, they propose selecting model

M_1 whenever a test for the hypothesis $\sigma^{-2}\theta' R'[R(X'X)^{-1}R']^{-1}R\theta \leq 1$ does not result in rejection, and selecting M_2 otherwise. It turns out that the appropriate test statistic is again the F statistic, but with a critical value that is chosen from a noncentral F distribution.

It is important to point out that the PMSE (aka "pretest" estimator) for θ resulting from first selecting model $\widehat{M}$ by some of the testing procedures described above and then estimating the parameters in model $\widehat{M}$ by least squares is neither the restricted nor the unrestricted least squares estimator, but a *random* convex combination of both; cf. (2). In particular, while it is true that the mean squared error of the restricted least squares estimator (corresponding to M_1) is smaller than the mean squared error of the unrestricted least squares estimator (corresponding to M_2) whenever model M_1 is true (and more generally, as long as $\sigma^{-2}\theta' R'[R(X'X)^{-1}R']^{-1}R\theta \leq 1$ holds), the PMSE need not (and will not) have a mean squared error equal to the better of the mean squared errors of the restricted and unrestricted estimators, but will be larger. Hence, if keeping the mean squared error of the PMSE small is the ultimate goal, one should set the significance level for the test underlying the model selection procedure such that the overshoot over the better of the mean squared errors of the restricted and unrestricted estimators does not exceed a prescribed "tolerance level." This has been investigated by Kennedy and Bancroft (1971), Sawa and Hiromatsu (1973), Brook (1976), Toyoda and Wallace (1976), Droge (1993), and Droge and Georg (1995); see also Section 10 in Amemiya (1980).

1.2.2 Procedures based on model selection criteria

If the ultimate goal of model selection is to find a model that gives rise to parameter estimators or predictors with small mean squared error (or some other risk measure) it seems to be natural to approach the model selection problem in a way that is geared towards this aim. The approach of Toro-Vizcarrondo and Wallace (1968) mentioned above combines the testing approach with such a risk-oriented approach. Alternatively, one can try to estimate the model associated with the smallest risk. [Whether or not the ensuing PMSE then actually has small risk is another matter; see the discussion further below.] To fix ideas consider the standard linear regression model (1) with errors that have mean zero and variance-covariance matrix $\sigma^2 I_n$. For any model $M \in \mathcal{M}_{all}$ let $\widehat{\theta}(M)$ denote the (restricted) least squares estimator computed under the zero-restrictions defining M. The mean squared error of $X\widehat{\theta}(M)$ is then given by

$$\begin{aligned} \mathrm{MSE}_{n,\theta}(M) &= \mathbb{E}_{n,\theta}\left\|X\widehat{\theta}(M) - X\theta\right\|^2 = \mathbb{E}_{n,\theta}\left\|P_M Y - X\theta\right\|^2 \\ &= \sigma^2 \mathrm{tr}(P_M) + \theta' X'(I - P_M)X\theta \\ &= \sigma^2 k_M + \theta' X'(I - P_M)X\theta, \end{aligned} \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm, P_M denotes projection on the column space spanned by the regressors active in M, and k_M denotes the number of these regressors. Ideally, we would like to use that model M that minimizes the risk (3), i.e., the model that has mean squared error equal to

$$\min_{M \in \mathcal{M}} \mathrm{MSE}_{n,\theta}(M), \tag{4}$$

where $\mathcal{M}$ is the set of candidate models specified by the researcher. The expression in (4) is sometimes called the "risk target" and it depends on the unknown parameters θ and σ^2 as well as on the set of candidate models $\mathcal{M}$ (and on X). However, since (3) (and (4)) are unobservable, it is not feasible to use the risk-minimizing model. An obvious idea is then to estimate (3) for every $M \in \mathcal{M}$ and to find the model that minimizes this estimator of the risk (sometimes called the "empirical risk"). An unbiased estimator for (3) is easily found as follows. Let M_{full} denote model (1), i.e., the model containing all K regressors, and let $\widehat{\theta}$ be shorthand for $\widehat{\theta}(M_{full})$. Then

$$\begin{aligned}\mathbb{E}_{n,\theta}\left(\widehat{\theta}' X'(I - P_M)X\widehat{\theta}\right) &= \mathbb{E}_{n,\theta}\left(Y' P_{M_{full}}(I - P_M)P_{M_{full}} Y\right) \\ &= \mathbb{E}_{n,\theta}\left(Y'(P_{M_{full}} - P_M)Y\right) \\ &= \sigma^2(K - k_M) + \theta' X'(I - P_M)X\theta.\end{aligned}$$

Since σ^2 can easily be estimated unbiasedly by $\widehat{\sigma}^2 = \widehat{\sigma}^2(M_{full}) = (n - K)^{-1} Y'(I - P_{M_{full}})Y$, an unbiased estimator for $\mathrm{MSE}_{n,\theta}(M)$ is found to be

$$\mathsf{MC}_n(M) = \widehat{\theta}' X'(I - P_M)X\widehat{\theta} + 2k_M\widehat{\sigma}^2 - K\widehat{\sigma}^2. \tag{5}$$

Noting that $X\widehat{\theta}$ equals $P_{M_{full}} Y$, we can rewrite (5) as

$$\mathsf{MC}_n(M) = \mathrm{RSS}(M) + 2k_M\widehat{\sigma}^2 - n\widehat{\sigma}^2, \tag{6}$$

where $\mathrm{RSS}(M) = Y'(I - P_M)Y$. After division by $\widehat{\sigma}^2$, this is known as Mallows's C_p, introduced in 1964; see Mallows (1965, 1973). The model selection procedure based on Mallows's C_p now returns that model $\widehat{M}$ which minimizes (6) over the set $\mathcal{M}$. It should be mentioned that Mallows did not advocate the minimum C_p strategy just described, but voiced concern about this use of C_p (Mallows (1965, 1973, 1995)).

It is important to note that the PMSE $\widetilde{\theta}$ for θ obtained via selection of the model minimizing (6) is a compound procedure, and is *not* identical to any of the least squares estimators $\widehat{\theta}(M)$ obtained from the models $M \in \mathcal{M}$; as pointed out before in (2), it rather is a *random* convex combination of these estimators. As a consequence, despite the construction of $\widehat{M}$ as a minimizer of an empirical version of the risk of the least squares estimators associated with the models M, it does *not* follow that the mean squared error of $\widetilde{\theta}$ is equal to (or close to) the risk target (4). In fact, it can overshoot the

risk target considerably. This comment applies mutatis mutandis also to the model selection procedures discussed below and we shall return to this issue also in Section 2.2.

A related criterion is the so-called final prediction error (FPE), which has become well known through the work of Akaike (1969, 1970), set in the context of selecting the order of autoregressive models. The same criterion was actually introduced earlier by Davisson (1965) also in a time series context, and was—according to Hocking (1976)—discussed by Mallows (1967) in a regression context. In the present context of a linear regression model it amounts to selecting the model M that minimizes

$$\mathsf{FPE}_n(M) = \mathrm{RSS}(M)(n - k_M)^{-1}(1 + k_M/n). \tag{7}$$

The derivation of the FPE is somewhat similar in spirit to the derivation of Mallows's C_p: Suppose that now the mean squared error of prediction

$$\begin{aligned} \mathrm{MSEP}_{n,\theta}(M) &= \mathbb{E}_{n,\theta} \left\| Y^* - X\widehat{\theta}(M) \right\|^2 \\ &= \sigma^2(n + k_M) + \theta' X'(I - P_M)X\theta \end{aligned} \tag{8}$$

is the quantity of interest, where $Y^* = X\theta + u^*$, with u^* having the same distribution as u, but is independent of u. [Note that in the linear regression model considered here $\mathrm{MSE}_{n,\theta}(M)$ and $\mathrm{MSEP}_{n,\theta}(M)$ only differ by the additive term $\sigma^2 n$; hence, this difference is immaterial and we have switched to $\mathrm{MSEP}_{n,\theta}(M)$ only to be in line with the literature.] For models M that are correct, the second term in (8) vanishes and—transposing Akaike's (1970) argument in the autoregressive case to the case of linear regression—it is proposed to estimate the unknown variance σ^2 in the first term by $\widehat{\sigma}^2(M) = (n - k_M)^{-1}\mathrm{RSS}(M)$. Upon division by n, this gives (7). Hence, $n\mathsf{FPE}_n(M)$ is an unbiased estimator for (8) *provided* the model M is correct. For incorrect models M this is not necessarily so, but it is suggested in Akaike (1969, 1970) that then the misspecification bias will make $\widehat{\sigma}^2(M)$ large, obviating the need to take care of the bias term $\theta' X'(I - P_M)X\theta$. While this is true for fixed θ and large n, ignoring the bias term seems to be an unsatisfactory aspect of the derivation of the FPE. [Note also that if one were to estimate σ^2 by $\widehat{\sigma}^2 = \widehat{\sigma}^2(M_{full})$ rather than $\widehat{\sigma}^2(M)$ in the above derivation, one would end up with the absurd criterion $\widehat{\sigma}^2(1 + k_M/n)$.]

Akaike's (1973) model selection criterion AIC is derived by similar means and—in contrast to Mallows's C_p or the FPE, which are limited to linear (auto)regressions—is applicable in general parametric models. The risk measure used here is not the mean squared error of prediction but the expected Kullback–Leibler discrepancy between $\mathbb{P}_{\widehat{\eta}(M)}$ and the true distribution of Y, where $\widehat{\eta}(M)$ denotes the maximum likelihood estimator based on model M. Akaike (1973) proposed an estimator for the Kullback–Leibler discrepancy that is approximately unbiased *provided* that the model M is a correct model. This estimator is given by $(n/2)\mathsf{AIC}_n(M)$, where

$$\mathsf{AIC}_n(M) = -2n^{-1} \log L_{n,M}(Y, \widehat{\eta}(M)) + 2\#M/n \ , \tag{9}$$

$L_{n,M}$ denotes the likelihood function corresponding to model M, and $\#M$ denotes the number of parameters in M. [The analysis in Akaike (1973) is restricted to independent and identically distributed data, but can be extended to more general settings; see, e.g., Findley (1985) for a treatment in the context of linear time series models, and Findley and Wei (2002) for vector autoregressive models.] The minimum AIC procedure now consists of selecting that model $\widehat{M}$ that minimizes AIC_n over the set $\mathcal{M}$. For the linear regression model (1) with errors $u \sim N(0, \sigma^2 I_n)$, σ^2 unknown, the criterion AIC_n reduces—up to an irrelevant additive constant—to

$$\mathsf{AIC}_n(M) = \log(\mathrm{RSS}(M)/n) + 2k_M/n. \tag{10}$$

If the error variance σ^2 is known, $\mathsf{AIC}_n(M)$ is—again up to an irrelevant additive constant—equal to $\mathsf{MC}_n(M)$ with $\widehat{\sigma}^2$ replaced by σ^2. For a very readable account of the derivation of the criteria discussed so far see Amemiya (1980).

A different approach to model selection, which is Bayesian in nature, was taken by Schwarz (1978). Given priors on the parameters in each model M and prior probabilities for each model (i.e., a prior on $\mathcal{M}$), one can compute the posterior probability for each model M given the data and one would then choose the model with the highest posterior probability. Schwarz (1978) showed that the leading terms in the posterior probabilities do not depend on the specific prior employed: He showed that the negative of the log posterior probabilities can—for large sample sizes—be approximated by $(n/2)\mathsf{BIC}_n(M)$, where

$$\mathsf{BIC}_n(M) = -2n^{-1} \log L_{n,M}(Y, \widehat{\eta}(M)) + \#M(\log n)/n. \tag{11}$$

The minimum Bayesian information criterion (BIC) procedure then selects the model $\widehat{M}$ that minimizes $\mathsf{BIC}_n(M)$ over $\mathcal{M}$.

Variants of the procedures A variant of the FPE, studied in Bhansali and Downham (1977), is FPE_α which reduces to the FPE for $\alpha = 2$; see also Shibata (1984). Shibata (1986b) and Venter and Steele (1992) discussed ways of choosing α such that the maximal (regret) risk of the ensuing PMSE is controlled; cf. also Foster and George (1994). Variants of the AIC/BIC obtained by replacing the $\log n$ term in (11) by some other function of sample size have been studied, e.g., in Hannan and Quinn (1979), Pötscher (1989), Rao and Wu (1989), and Shao (1997); cf. also Section 2.1. As noted, the AIC is an asymptotically unbiased estimator of the Kullback–Leibler discrepancy if the model M is correct. A finite-sample bias correction for correct models M was provided by Sugiura (1978) and subsequently by Hurvich and Tsai (1989) and leads to the corrected AIC (AICC), which in the Gaussian linear regression context takes the form

$$\mathsf{AICC}_n(M) = \log(\mathrm{RSS}(M)/n) + 2(k_M + 1)/(n - k_M - 2).$$

The derivation of the AIC or AICC from asymptotically unbiased estimators for the Kullback–Leibler discrepancy is based on the assumption that the models M are correct models. For bias corrections allowing for M to be incorrect and for resulting model selection criteria, see Reschenhofer (1999) and references therein. Takeuchi's (1976) information criterion (TIC) should also be mentioned here. It is an approximately unbiased estimator of Kullback–Leibler discrepancy also for incorrect models. However, the TIC requires consistent estimators for the expectations of the Hessian of the log-likelihood as well as of the outer product of the score, where the expectation is taken under the true distribution. A possibility to implement this is to use bootstrap methods; see Shibata (1997). The derivations underlying the AIC or the TIC assume that the models are estimated by maximum likelihood. Konishi and Kitagawa (1996) introduced a model selection criterion GIC that allows for estimation procedures other than maximum likelihood. See also the recent book by Konishi and Kitagawa (2008). The derivation of the FPE in autoregressive models is based on the one-step-ahead prediction error. Model selection criteria that focus on multi-step-ahead predictors can similarly be derived and are treated in Findley (1991), Bhansali (1999), and Ing (2004).

Other model selection criteria Myriads of model selection criteria have been proposed in the literature and it is impossible to review them all. Here we just want to mention some of the more prominent criteria not yet discussed. Theil (1961) was perhaps the first to suggest using the adjusted R^2 as a model selection criterion. Maximization of the adjusted R^2 amounts to minimization (with respect to M) of

$$(n - k_M)^{-1}\mathrm{RSS}(M).$$

Another early criterion that was apparently suggested by Tukey is given by

$$S_n(M) = ((n - k_M)(n - k_M - 1))^{-1}\mathrm{RSS}(M). \tag{12}$$

It is—similarly to Mallows's C_p—obtained from an unbiased estimator of the out-of-sample mean squared error of prediction in a linear regression model, where now the vector of regressors is assumed to be independent and identically normally distributed with mean zero and the expectation defining the mean squared error of prediction is also taken over the regressors (in the observation as well as in the prediction period). This criterion is further discussed in Thompson (1978a, b), Breiman and Freedman (1983), and Leeb (2006b). Cross-validation provides another method for model selection (Allen (1974); Stone (1974); Shao (1993); Zhang (1993a); Rao and Wu (2005)). Generalized cross-validation was introduced by Craven and Wahba (1979) and in the linear regression context of Example 1 amounts to minimizing

$$\mathrm{GCV}_n(M) = (n - k_M)^{-2}\mathrm{RSS}(M).$$

Note the close relationship with Tukey's $S_n(M)$. Also in the context of a standard linear regression model with nonstochastic regressors, Foster and George (1994) introduced the so-called risk inflation criterion based on considering the minimization of the maximal inflation of the risk of the PMSE over the infeasible "estimator" that makes use of the knowledge of the (minimal) true model. This criterion is given by

$$\mathsf{RIC}_n(M) = \mathrm{RSS}(M) + 2k_M \log(K)\widehat{\sigma}^2.$$

See also George and Foster (2000). On the basis of considerations of the code length necessary to encode the given data by encoding the fitted candidate models, Rissanen (1978, 1983, 1986a, b, 1987) introduced the minimum description length (MDL) criterion and the closely related predictive least squares (PLS) criterion; cf. also the review article by Hansen and Yu (2001) as well as Rissanen (1989). These criteria are also closely connected to Wei's (1992) Fisher information criterion (FIC). For more on this criterion see Chan (2008). Finally, if prediction at a value $x_{f\cdot}$ of the regressor vector different from the values in the sample is of interest and if the steps in the derivation of Mallows's C_p are repeated for this target, one ends up with a criterion introduced in Allen (1971). This criterion depends on the chosen $x_{f\cdot}$ and hence is an early precursor to the so-called focused information criterion of Claeskens and Hjort (2003). For a discussion of further model selection criteria see Rao and Wu (2001).

Relationships between criteria For many model selection problems such as, e.g., variable selection in linear regression or order selection for autoregressive processes the criteria AIC, AICC, FPE, Mallows's C_p, Tukey's S_n, cross-validation, as well as generalized cross-validation are "asymptotically equivalent" (Stone (1977); Shibata (1989)). These asymptotic equivalence results typically hold only for quite "small" families $\mathcal{M}$ of candidate models (e.g., for fixed finite families); in particular, k_M typically has to be small compared with n for the asymptotic equivalence results to bear on the finite-sample behavior. If k_M is not small relative to n, the asymptotic equivalence does not apply and these criteria can behave very differently. For more discussion on the relationship between various criteria see, e.g., Söderström (1977), Amemiya (1980), Teräsvirta and Mellin (1986), and Leeb (2006b).

A comment Criteria like Mallows's C_p, the AIC, the FPE, and so on have been derived as (asymptotically) unbiased estimators for mean squared error (of prediction) or Kullback–Leibler discrepancy for certain estimation problems. Often these criteria are also used in contexts where they are not necessarily (asymptotically) unbiased estimators (e.g., in a pseudo-likelihood context), or where no formal proof for the unbiasedness property has been provided. For example, for Gaussian autoregressive models the AIC is (approximately) unbiased (Findley and Wei (2002)) and takes the form $\log \widehat{\sigma}^2(k) + 2k/n$, where $\widehat{\sigma}^2(k)$ is the usual residual variance estimator from

an AR(k) fit. This latter formula is, however, routinely also used for model selection in autoregressive models with non-Gaussian (even heavy-tailed) errors without further justification. Another example is model selection in GARCH models, where procedures like minimum AIC are routinely applied, but formal justifications do not seem to be available.

Relationship between model selection criteria and hypothesis tests The model selection procedures described in Section 1.2.1 and in the present section are closely related. First observe that in a setting with only two nested models, i.e., $M_1 \subseteq M_2$, the minimum AIC procedure picks model M_2 if and only if the usual likelihood ratio test statistic of the hypothesis M_1 versus M_2 exceeds the critical value $2(k_{M_2} - k_{M_1})$. In general, the model M selected by minimum AIC is characterized by the property that the likelihood ratio test statistic for testing M against any other model $M' \in \mathcal{M}$ (nesting M or not) exceeds the respective critical value $2(k_{M'} - k_M)$. For more discussion, see Söderström (1977), Amemiya (1980), Teräsvirta and Mellin (1986), and Section 4 in Pötscher (1991).

2 Properties of Model Selection Procedures and of Post–Model–Selection Estimators

We now turn to the statistical properties of model selection procedures and their associated PMSEs. In particular, questions like consistency/inconsistency of the model selection procedure, risk properties, as well as distributional properties of the associated PMSE are discussed. In this section we concentrate on the case where the set $\mathcal{M}$ of candidate models contains a correct model; furthermore, $\mathcal{M}$ is here typically assumed to be finite, although some of the results mentioned in this section also hold if $\mathcal{M}$ expands suitably with sample size or is infinite. The case of model selection from a set of models that are potentially only approximations to the data-generating mechanism is discussed in Section 3.

2.1 Selection probabilities and consistency

The focus in this subsection is on the model selection procedure $\widehat{M}$ viewed as an estimator for the minimal true model (given that it exists). For definiteness of discussion, consider the linear regression model as in Example 1 with an $N(0, \sigma^2 I_n)$-distributed error term. Assume for simplicity of presentation further that the set $\mathcal{M} \subseteq \mathcal{M}_{all}$ of candidate models contains the full model M_{full} and is stable with respect to intersections, meaning that with M and M' belonging to $\mathcal{M}$, also $M \cap M'$ belongs to $\mathcal{M}$. [This is, e.g., the

case for $\mathcal{M} = \mathcal{M}_{all}$ or $\mathcal{M} = \mathcal{M}_{nested}$.] Under this condition, for each value of the parameter $\theta \in \mathbb{R}^K$ there exists a minimal true model $M_0 = M_0(\theta)$ given by

$$M_0 = \bigcap_{\theta \in M \in \mathcal{M}} M.$$

[If $\mathcal{M} = \mathcal{M}_{all}$, then M_0 is given by the set of all parameters θ^* that have $\theta_i^* = 0$ whenever $\theta_i = 0$. If $\mathcal{M} = \mathcal{M}_{nested}$, then M_0 is given by the set of all parameters θ^* that have $\theta_i^* = 0$ for all $i > p_0(\theta)$, where $p_0(\theta)$ is the largest index such that $\theta_{p_0(\theta)} \neq 0$ (and $p_0(\theta) = 0$ if $\theta = 0$).] The quality of $\widehat{M}$ as an estimator for M_0 can be judged in terms of the "overestimation" and "underestimation" probabilities, respectively, i.e., in terms of the probabilities of the events

$$\{\widehat{M} \neq M_0, \widehat{M} \supseteq M_0\} \tag{13}$$

and

$$\{\widehat{M} \not\supseteq M_0\}. \tag{14}$$

Note that (13) represents the case where a correct model containing superfluous regressors is selected, whereas (14) describes the case where an incorrect model is selected.

A model selection procedure is *consistent* if the probabilities of overestimation and underestimation converge to zero, i.e., if

$$\lim_{n \to \infty} P_{n,\theta}(\widehat{M} = M_0) = 1 \tag{15}$$

for every $\theta \in \mathbb{R}^K$. If

$$\lim_{n \to \infty} P_{n,\theta}(\widehat{M} \not\supseteq M_0) = 0 \tag{16}$$

for every $\theta \in \mathbb{R}^K$, but $\widehat{M}$ is not consistent, we say that $\widehat{M}$ is *conservative*. We note that any reasonable model selection procedure will satisfy (16). Suppose that in the context of the linear regression model considered here the regressors also satisfy the "asymptotic stationarity" condition $X'X/n \to Q$, where Q is a positive-definite matrix. If $\widehat{M}$ is then obtained through minimization of a criterion of the form

$$\mathsf{CRIT}(M) = \log(\mathrm{RSS}(M)/n) + k_M C_n/n, \tag{17}$$

it is well known that $\widehat{M}$ is consistent if the penalty satisfies $C_n/n \to 0$ and $C_n \to \infty$ as $n \to \infty$; and it is conservative if C_n is bounded (e.g., Geweke and Meese (1981)). In particular, it follows that the minimum BIC procedure (i.e., $C_n = \log n$) is consistent, whereas the minimum AIC procedure (i.e., $C_n = 2$) is conservative. [That the FPE is conservative was already noted in Akaike (1970) in the context of selecting the order of stationary autoregressions.] If the asymptotic stationarity condition $X'X/n \to Q$ does not hold, the conditions on C_n for consistency/conservatism change and are related to

the rate of increase of the largest and smallest eigenvalues of $X'X$ (Pötscher (1989)). We note that these results are not tied to the normality assumption or the assumption of independent and identically distributed errors made here for the sake of simplicity, but hold under more general assumptions. For further consistency results in the context of linear regression models, see Nishii (1984), An and Gu (1985), Rao and Wu (1989), and Zheng and Loh (1995, 1997); Chen and Ni (1989) provide consistency results for linear regression models with short-memory time series errors, while Ing and Wei (2006) allow also for long-memory errors; see also Hidalgo (2002). Consistency results for model selection based on (17) or on closely related criteria, but for model classes other than linear regression, can be found in Hannan and Quinn (1979) and Quinn (1980) for stationary autoregressions and in Hannan (1980, 1981) for stationary autoregressive moving average (ARMA) models (see also An and Chen 1986 and Chapter 5 of Hannan and Deistler (1988) for more discussion and references); in Paulsen (1984), Tsay (1984), Pötscher (1989), and Wei (1992) for nonstationary autoregressions (the last two of these papers considering also more general classes of stochastic regression models); in Knight (1989) for infinite variance autoregressions; in Kohn (1983) and Nishii (1988) for general parametric models; and in Haughton (1991) for nonlinear regression models. Similar results hold for criteria like FPE_α if α is made dependent on sample size in an appropriate manner and are discussed in several of the references just given. Consistency results for the PLS criterion can be found, e.g., in Rissanen (1986b), Hemerly and Davis (1989), and Wei (1992). The papers on consistency mentioned so far consider a finite set of candidate models (which in some results is allowed to expand with sample size, typically slowly). In the context of order selection of stationary ARMA models, Pötscher (1990) discussed a modification of BIC-like procedures and established consistency without any restriction on the size of the set of candidate models, i.e., the result applies even for $\mathcal{M}$ being the (infinite) set of *all* ARMA models; see also Pötscher and Srinivasan (1994). We are not aware of any published formal results establishing consistency of model selection procedures in GARCH models, although such results are certainly possible. Francq et al. (2001) established that the AIC and related criteria are conservative procedures in the context of autoregressive conditional heteroscedasticity (ARCH) models.

Model selection procedures based on tests are typically consistent if the critical values employed by the tests are chosen in such a way that they diverge to infinity at an appropriate rate. Otherwise the procedures are typically conservative. Consistency results of this sort are provided in Pötscher (1983) in the context of selecting the order of ARMA models and in Bauer et al. (1988) for a "thresholding" procedure in a general (semi)parametric model. For a follow-up on the latter paper, see Bunea et al. (2006).

The limits of the model selection probabilities in the case of conservative model selection procedures were studied in Shibata (1976), Bhansali and Downham (1977), Hannan (1980), Geweke and Meese (1981), Sakai (1981),

Tsay (1984), and Quinn (1988). Further studies of the overestimation and underestimation probabilities in various settings and for various model selection procedures can be found in Zhang (1993b), Shao (1998), Guyon and Yao (1999), and Keribin and Haughton (2003).

Since it seems to be rarely the case that estimation of M_0 is the ultimate goal of the analysis, the consistency property of $\widehat{M}$ may not be overly important. In fact, as we shall see in the next subsection, consistency of $\widehat{M}$ has detrimental effects on the risk properties of the associated PMSE. This may seem to be counterintuitive at first sight and is related to the fact that the consistency property (15) does not hold uniformly with respect to the parameter θ (Remark 4.4 in Leeb and Pötscher (2005)). Furthermore, in a situation where none of the models in the class $\mathcal{M}$ are correct (see Section 3), that is, if only "approximate" models are fitted, the concept of consistency becomes irrelevant.

2.2 Risk properties of post-model-selection estimators

As already noted in Sects. 1.2.1 and 1.2.2 it is important to realize that—despite the ideas underlying the construction of PMSEs—a PMSE does not come with an automatic optimality property; in particular, its risk is by no means guaranteed to equal the risk target (4). For example, while $\mathsf{MC}_n(M)$ given by (5) is an unbiased estimator of the mean squared error $\mathrm{MSE}_{n,\theta}(M)$ of the estimator $\widehat{\theta}(M)$ based on model M, minimizing $\mathsf{MC}_n(M)$ gives rise to a *random* $\widehat{M}$ and an associated PMSE $\widetilde{\theta}$ that is a *random* convex combination of the estimators $\widehat{\theta}(M)$ based on the various models $M \in \mathcal{M}$. As a consequence, the mean squared error of $\widetilde{\theta}$ is no longer described by any of the quantities $\mathrm{MSE}_{n,\theta}(M)$ (or by the risk target (4) for that matter), since $\widetilde{\theta}$ falls outside the class $\{\widehat{\theta}(M) : M \in \mathcal{M}\}$.

The risk properties of PMSEs have been studied for model selection procedures based on test procedures in considerable detail; see Judge and Bock (1978), Giles and Giles (1993), Magnus (1999), and Danilov and Magnus (2004). For procedures based on model selection criteria, investigations into the risk properties of PMSEs can be found in Mallows (1973), Hosoya (1984), Nishii (1984), Shibata (1984, 1986b, 1989), Venter and Steele (1992), and Foster and George (1994). A basic feature of risk functions of PMSEs that emerges from these studies is best understood in the context of the simple normal linear regression model (Example 1) when selection is only between two models M_1 and $M_2 = M_{full}$, where M_1 is obtained from M_2 by restricting the $k_2 \times 1$ subvector θ_2 of the $(k_1 + k_2) \times 1$ parameter vector $\theta = (\theta_1', \theta_2')'$ to zero. Recall that in this example the quadratic risk $\mathrm{MSE}_{n,\theta}(\widehat{\theta}(M_1))$ is an unbounded quadratic function of θ_2, which achieves its minimal value $\sigma^2 k_1$ on the set $\{\theta : \theta_2 = 0\}$, i.e., when model M_1 holds, whereas the

quadratic risk $\mathrm{MSE}_{n,\theta}(\widehat{\theta}(M_2))$ is constant and equals $\sigma^2(k_1 + k_2)$. Hence, $\mathrm{MSE}_{n,\theta}(\widehat{\theta}(M_1)) < \mathrm{MSE}_{n,\theta}(\widehat{\theta}(M_2))$ whenever $\theta_2 = 0$, and this inequality persists for $\theta_2 \neq 0$ sufficiently small (by continuity of the mean squared error). However, as θ_2 moves away from the origin, eventually the inequality will be reversed. Now, the risk $\mathrm{MSE}_{n,\theta}(\widetilde{\theta})$ of the PMSE $\widetilde{\theta}$ will typically also be less than the risk of $\widehat{\theta}(M_2)$ for parameter values which have $\|\theta_2\|$ sufficiently close to zero, but it will rise *above* the risk of $\widehat{\theta}(M_2)$ as $\|\theta_2\|$ becomes larger; eventually the risk of $\widetilde{\theta}$ will attain its maximum and then gradually approach the risk of $\widehat{\theta}(M_2)$ from above as $\|\theta_2\|$ increases further and approaches infinity. As stressed by Magnus (1999), for many PMSEs there will be even regions in the parameter space (for intermediate values of $\|\theta_2\|$) where the risk of the PMSE will actually be larger than the larger of the risks of $\widehat{\theta}(M_1)$ and $\widehat{\theta}(M_2)$. Figure 5 in Leeb and Pötscher (2005) gives a representation of the typical risk behavior of a PMSE.

2.2.1 Limiting risk in the case of consistent model selection

Continuing the previous example, suppose $\widehat{M}$ is a consistent model selection procedure for M_0, i.e., $\widehat{M}$ satisfies (15), where the minimal true model M_0 is here given by

$$M_0 = \begin{cases} M_1 \text{ if } \theta_2 = 0 \\ M_2 \text{ if } \theta_2 \neq 0 \end{cases}.$$

Assume also that $X'X/n \to Q > 0$ for $n \to \infty$. Then $P_{n,\theta}(\widehat{M} = M_1)$ will typically go to zero exponentially fast for $\theta_2 \neq 0$ (Nishii (1984)). It is then easy to see that

$$\lim_{n\to\infty} \mathrm{MSE}_{n,\theta}(\widetilde{\theta}) = \begin{cases} \sigma^2 k_1 & \text{if } \theta_2 = 0 \\ \sigma^2(k_1 + k_2) & \text{if } \theta_2 \neq 0 \end{cases}. \tag{18}$$

That is, for each *fixed* θ the limiting risk of the PMSE coincides with the (limiting) risk of the restricted estimator $\widehat{\theta}(M_1)$ if model M_1 obtains, and with the (limiting) risk of the unrestricted estimator $\widehat{\theta}(M_2)$ if model M_2 is the minimal true model. Put yet another way, the limiting risk of the PMSE coincides with the (limiting) risk of the "oracle," i.e., of the infeasible "estimator" based on the minimal true model M_0. This *seems* to tell us that in large samples consistent model selection typically results in a PMSE that has approximately the same risk behavior as the *infeasible* procedure that uses $\widehat{\theta}(M_1)$ if $\theta_2 = 0$ and uses $\widehat{\theta}(M_2)$ if $\theta_2 \neq 0$. Note that this procedure is infeasible (and hence is sometimes dubbed an "oracle"), since it uses knowledge of whether $\theta_2 = 0$ or not. The above observation that in a "pointwise" asymptotic analysis (that is, in an asymptotic analysis that holds the true parameter fixed while letting sample size increase) a consistent model selec-

tion procedure typically has no effect on the limiting risk of the PMSE was made in Nishii (1984); see also the discussion of the so-called oracle property in Section 2.3. Unfortunately, this result—while mathematically correct—is a statistical fallacy and does not even approximately reflect the risk properties at any given sample size, regardless of how large: It can be shown that—despite (18)—the worst-case risk of *any* PMSE based on a consistent model selection procedure diverges to infinity, i.e.,

$$\lim_{n\to\infty} \sup_{\theta} \mathrm{MSE}_{n,\theta}(\widetilde{\theta}) = \infty \tag{19}$$

holds. Hence, in terms of worst-case risk a PMSE based on a consistent model selection procedure is much worse than, e.g., the least squares estimator based on the overall model (which has constant risk $\sigma^2(k_1 + k_2)$), or than a PMSE based on a conservative procedure (which typically has bounded worst-case risk, cf. (20)). This phenomenon, which is in striking contrast to (18), has been observed at different levels of generality by Hosoya (1984), Shibata (1986b), Foster and George (1994), Leeb and Pötscher (2005, 2008a), and Yang (2005, 2007). As shown in Leeb and Pötscher (2008a), the unboundedness phenomenon (19) also applies to so-called sparse estimators as considered, e.g., in Fan and Li (2001); cf. Section 4. We note that the finite-sample behavior of the risk function of the PMSE, which gets lost in a pointwise asymptotic analysis, can be captured in an asymptotic analysis that makes the true parameter dependent on sample size; see Leeb and Pötscher (2005).

2.2.2 Limiting risk in the case of conservative model selection

Results on the limiting risk in this case were provided in Hosoya (1984), Nishii (1984), Shibata (1984), and Zhang (1992). For conservative model selection procedures the limiting risk of the associated PMSE does not satisfy (18). In fact, in this case it can be shown that the limiting risk of a PMSE is typically larger than the limiting risk of the corresponding oracle (i.e., the infeasible "estimator" based on the minimal true model M_0), except if the minimal true model is the overall model. Hence, for conservative model selection procedures a pointwise asymptotic analysis already reveals some of the effects of model selection on the risk of the PMSE, although the full effect is again only seen in an asymptotic analysis that makes the true parameter dependent on sample size (or in a finite-sample analysis, of course); see Leeb and Pötscher (2005) for an extensive discussion. In contrast to PMSEs based on consistent model selection procedures, the worst-case risk of a PMSE based on a conservative procedure is typically bounded as sample size goes to infinity, i.e.,

$$\lim_{n\to\infty} \sup_{\theta} \mathrm{MSE}_{n,\theta}(\widetilde{\theta}) < \infty \tag{20}$$

typically holds.

2.2.3 Admissibility results

Admissibility or inadmissibility of PMSEs in various classes of estimators was discussed in Sclove et al. (1972), Stone (1981, 1982), Takada (1982), and Kempthorne (1984).

2.2.4 Consistency of post-model-selection estimators

A PMSE is typically consistent regardless of whether the model selection procedure is consistent or conservative; this follows, e.g., from Lemma 2 in Pötscher (1991). In fact, PMSEs will often be even uniformly consistent; cf. Propositions A.9 and B.1 in Leeb and Pötscher (2005) and Theorem 2 in Pötscher and Leeb (2007).

2.3 Distributional properties of post-model-selection estimators

As noted in Section 1.1, a PMSE $\widetilde{\eta}$ is a *random* convex combination of the estimators $\widehat{\eta}(M)$ computed on the basis of model M. As a consequence, the distribution of $\widetilde{\eta}$ will typically be more complex than the distribution of $\widehat{\eta}(M)$, which often will be asymptotically normal. Sen (1979) derived the asymptotic distribution of $n^{1/2}(\widetilde{\eta} - \eta)$ in a maximum likelihood framework for independent and identically distributed data when the model selection procedure consists in choosing from two nested models $M_1 \subseteq M_2$ on the basis of the likelihood ratio test. Pötscher (1991) obtained the asymptotic distribution for the case when model selection is from a nested family $M_1 \subseteq M_2 \subseteq \ldots \subseteq M_P$ and is based on a general-to-specific hypothesis testing scheme; the framework in Pötscher (1991) is also more general than the one in Sen (1979) in that it allows for dependent data and M-estimators other than maximum likelihood. Furthermore, Pötscher (1991) derived not only the unconditional, but also the conditional asymptotic distribution of $n^{1/2}(\widetilde{\eta} - \eta)$. Here the conditioning is on the event of having chosen a particular model. See Pötscher and Novak (1998) for further results and a simulation study. In the same framework as in Pötscher (1991), but confining attention to the normal linear regression model, Leeb and Pötscher (2003) and Leeb (2005, 2006a) obtained the unconditional as well as the conditional finite-sample distribution of $n^{1/2}(\widetilde{\eta} - \eta)$ (as well as their limits under local alternatives). From the above references it transpires that the asymptotic as well as the finite-sample distributions of $n^{1/2}(\widetilde{\eta} - \eta)$ are complicated and, in particular, are typically decidedly non-normal, e.g., they can be bimodal. Furthermore, these distributions depend on the unknown parameter η in a complicated way. As a consequence of these results, the usual confidence intervals naively applied to PMSEs do not have

correct coverage probability, not even asymptotically (Saleh and Sen (1983); Pötscher (1991); Zhang (1992); Kabaila (1998); Kabaila and Leeb (2006)). For further results on distributional properties of PMSEs based on conservative model selection procedures, see Sen and Saleh (1987), Dijkstra and Veldkamp (1988), Kabaila (1995), Pötscher (1995), Ahmed and Basu (2000), and Hjort and Claeskens (2003). Shen et al. (2004) in their Theorem 3 incorrectly claim that the asymptotic distribution of the PMSE based on, say, the AIC, is normal.

The results discussed in this subsection so far apply to *conservative* model selection procedures. It is important to note, however, that the finite-sample results in Leeb and Pötscher (2003) and Leeb (2005, 2006a) also apply to *consistent* model selection procedures based on general-to-specific testing (i.e., procedures where the critical values diverge to infinity at an appropriate rate with sample size), since for fixed sample size n it is irrelevant whether we view the critical values as being constant or as depending on n. Hence, the conclusions regarding the finite-sample distributions of PMSEs drawn in the previous paragraph carry over to the case of consistent model selection. When it comes to the *pointwise* asymptotic distribution of PMSEs based on *consistent* model selection procedures a difference arises: It is easy to see that the pointwise asymptotic distribution of $n^{1/2}(\widetilde{\eta}-\eta)$ is then typically normal and coincides with the (pointwise) asymptotic distribution of $n^{1/2}(\widehat{\eta}(M_0)-\eta)$, where $\widehat{\eta}(M_0)$ is the infeasible "oracle" that makes use of knowledge of the minimal true model M_0. This was noted in Hannan and Quinn (1979) and Lemma 1 in Pötscher (1991), who also issued a warning regarding the statistical interpretation of this result. Nevertheless, this property of PMSEs based on consistent model selection procedures has frequently—and incorrectly—been interpreted in the literature as saying that consistent model selection has no effect asymptotically on the distributional properties of the parameter estimator and that one can estimate θ asymptotically as efficient as if knowledge about the minimal true model were available: Two prominent examples are Bunea (2004) and Fan and Li (2001), who advertise this property of their estimators, the latter authors dubbing this property the "oracle property." However, this interpretation is a fallacy: The "oracle property" is essentially a reincarnation of the "superefficiency" phenomenon à la the "superefficiency" of Hodges's estimator, and does not reflect actual statistical performance. Other instances in the literature where this misleading interpretation has been reported include Geweke and Meese (1981), Lütkepohl (1990, p. 120), Hidalgo (2002), Hall and Peixe (2003), and Dufour et al. (2006). Mathematically speaking, the problem is that convergence of the finite-sample distributions to their asymptotic counterparts is highly nonuniform in the parameter, and that the "oracle property" results are only *pointwise* asymptotic results. This has already been noted by Shibata (1986a) and Kabaila (1995, 1996). While pointwise asymptotics are unable to capture the effects of consistent model selection, a more appropriate asymptotic analysis that allows the true parameter to depend on sample size very well reveals these effects;

see Leeb and Pötscher (2005) for an extensive discussion. The nonuniformity in the convergence of the finite-sample distributions is of course related to the unboundedness of the maximal risk, cf. (19), discussed in the previous subsection.

2.3.1 Estimation of the distribution of post-model-selection estimators

As discussed already, the finite-sample (as well as the asymptotic) distribution of PMSEs typically depends on unknown parameters in a complicated fashion. In order to be able to utilize such distributions for inference one has to estimate these distributions. While consistent estimators for the distribution of PMSEs can be constructed, it was shown in Leeb and Pötscher (2006b, 2008b) that such estimators are necessarily of low quality in the sense that no estimator can be uniformly consistent. Such "impossibility" results also arise for a large class of shrinkage-type estimators: see Leeb and Pötscher (2006a), Pötscher and Leeb (2007), and Pötscher and Schneider (2007). See also Section 2.3 in Leeb and Pötscher (2005) for a simple exposition of the issues involved here.

2.3.2 Confidence sets post model selection

Problems related to the construction of valid confidence intervals after model selection are discussed in Kabaila (1995, 1998), Pötscher (1995), Kabaila and Leeb (2006), Leeb (2007), and Pötscher (2007).

3 Model Selection in Large- or Infinite–Dimensional Models

In Section 2 we mainly concentrated on the case where there exists a true model in the set of candidate models $\mathcal{M}$ and where the cardinality of $\mathcal{M}$ is finite and independent of sample size n. It can, however, be argued that the need for model selection is particularly great when the dimension of the candidate models (e.g., number of potentially important explanatory variables) is large in relation to sample size and/or no model in $\mathcal{M}$ is correct (i.e., the models fitted to the data constitute only an approximation to the data-generating process). To analyze a scenario like this, it is often more appropriate to assume that the true data-generating process is infinite-dimensional, and that one tries to identify a "good" finite-dimensional model on the basis of the data.

Throughout this section, we consider the following "infinite-dimensional" extension of Example 1: In the setting of that example, assume that the number of regressors, i.e., K in (1), is infinite. To ensure that the response Y and its mean $X\theta$ are well defined, assume that θ as well as the row-vectors of X are square-summable, i.e., θ and $X'_{i,\cdot}$ are in l_2. Moreover, assume that the infinite-dimensional "matrices" (i.e., operators on l_2) $X'X/n$ converge in the operator norm to an invertible operator Q as $n \to \infty$. We also make the assumption that the vector of errors u is distributed as $N(0, \sigma^2 I_n)$, $0 < \sigma^2 < \infty$. Given a sample of size n, consider model selection from a family $\mathcal{M}_n$ of finite-dimensional candidate models. Throughout, we always assume that each model $M \in \mathcal{M}_n$ is such that the $n \times k_M$ matrix of those explanatory variables included in the model M has full column-rank k_M; we make this assumption for convenience, although it is not necessary for all the results discussed below. The collection of candidate models $\mathcal{M}_n$ considered at sample size n is assumed to be finite or countable, and it is allowed to depend on sample size satisfying $\mathcal{M}_n \subseteq \mathcal{M}_{n+1}$ (although this is again not necessary for all the results discussed below). This setting is sufficiently general to present the relevant results, while it is sufficiently simple to keep the notation and assumptions manageable. We refer to the literature for more general results.

One of the early analyses of model selection in infinite-dimensional models is Shibata (1980); see also Shibata (1981a, b). In essence, these papers establish that the PMSE based on the AIC (or the FPE) is pointwise asymptotically loss-efficient *provided* that the true data-generating process is infinite-dimensional (and that the collection of candidate models $\mathcal{M}_n$ increases appropriately with n). In the setting of (1) with $K = \infty$ as introduced above, define the loss $L_n(\theta, \bar{\theta})$ of an estimator $\bar{\theta}$ of θ (taking values in l_2) as

$$L_n(\theta, \bar{\theta}) = (\bar{\theta} - \theta)' \frac{X'X}{n} (\bar{\theta} - \theta),$$

and let $R_n(\theta, \bar{\theta}) = \mathbb{E}_{n,\theta}(L_n(\theta, \bar{\theta}))$ denote the corresponding risk. Given a model selection procedure $\widehat{M}$, Shibata (1981b) compared the loss of the PMSE $\widehat{\theta}(\widehat{M})$, i.e., $L_n(\theta, \widehat{\theta}(\widehat{M}))$ with the minimum of the losses of the least-squares estimators $\widehat{\theta}(M)$ corresponding to all the models M in $\mathcal{M}_n$, i.e., with $\inf_{M \in \mathcal{M}_n} L_n(\theta, \widehat{\theta}(M))$. If $\widehat{M}_{AIC}$ is chosen by the minimum AIC method, i.e., $\widehat{M}_{AIC}$ is a (measurable) minimizer of $\mathsf{AIC}_n(M)$ over $M \in \mathcal{M}_n$, then Shibata (1981b) showed that

$$\frac{L_n(\theta, \widehat{\theta}(\widehat{M}_{AIC}))}{\inf_{M \in \mathcal{M}_n} L_n(\theta, \widehat{\theta}(M))} \xrightarrow{p} 1 \tag{21}$$

provided that the true parameter θ is truly infinite dimensional (i.e., has infinitely many nonzero coordinates), provided that the candidate models are nested in the sense that $\mathcal{M}_n = \{M(p) : 0 \le p \le K_n\}$, and provided that the number of candidate models $K_n + 1$ satisfies $K_n \to \infty$ and $K_n = o(n)$.

[The model $M(p)$ here refers to the model containing the first p regressors; cf. Example 1.] Relation (21) continues to hold if in the denominator loss is replaced by risk. All these results carry over if, instead of the AIC, either Mallows's C_p or the FPE is used for model selection (cf. Theorems 2.1, 2.2, 3.1, and the discussion leading up to Assumption 1, as well as Section 5 in Shibata (1981b)). Shibata (1981b) pointed out that (21) does not hold for model selectors based on BIC-type model selection criteria. The results in Shibata (1981b) in fact allow for classes of candidate models more general than the nested case considered here, provided that the condition that θ is infinite-dimensional is replaced by a more complicated condition that, in essence, requires that the candidate models considered at sample size n do not fit the true data-generating process "too well"; see Assumptions 1 and 2 in Shibata (1981b). For order selection in autoregressive models approximating an infinite-order autoregressive data-generating process, results similar to those just presented were given by Shibata (1980); here models were evaluated in terms of their predictive performance out of sample, where the model selection and fitting step on the one hand and the prediction step on the other hand are based on two independent realizations of the same time series. Recently, Ing and Wei (2005) obtained parallel results for the case where one and the same realization of the process is used for the estimation, selection, and prediction step. Shibata (1981a) also considered selection of approximating autoregressive models where the models were evaluated by the performance of the corresponding estimate of the spectral density (at a fixed frequency). Pointwise asymptotic loss-efficiency results in the above sense were also established in Li (1987), Polyak and Tsybakov (1990), and Shao (1997) for a variety of other methods, including generalized cross-validation and cross-validation, and under somewhat different sets of assumptions. See also Breiman and Friedman (1983) for a similar result about the criterion (12) when models are evaluated by their predictive performance out of sample.

All the pointwise asymptotic loss-efficiency results for conservative model selection procedures like (21) mentioned above rely on the central assumption that the true data-generating process is "not too well approximated" by the finite-dimensional candidate models considered at sample size n as $n \to \infty$. In the simple setting considered in (21), this is guaranteed by assuming that θ is infinite-dimensional, and, in more general settings, by conditions like Assumption 2 in Shibata (1981b). If that central assumption is violated, statements like (21) will typically break down for conservative model selection procedures (like the AIC or the FPE). In fact, Shao (1997) considered a scenario where the BIC and related consistent model selection procedures are pointwise asymptotically loss-efficient when the true model is finite-dimensional. [This is in line with the discussion of the "oracle" phenomenon in Section 2.2 for the case of finitely many candidate models.] These findings suggest a dichotomy: If the true model is infinite-dimensional, conservative procedures like the AIC are pointwise asymptotically loss-efficient, while consistent procedures like the BIC are not; if the true model is finite-dimensional, the

situation is reversed (under appropriate assumptions). However, one should not read too much into these results for the following reasons: The true model may be infinite-dimensional, suggesting an advantage of the AIC or a related procedure over the BIC. At a given sample size, however, one of the finite-dimensional candidate models may provide a very good approximation to the true data-generating process, and hence the pointwise asymptotic loss-efficiency result favoring the AIC may not be relevant. Conversely, the true model may be finite-dimensional, suggesting an advantage of consistent procedures like the BIC, but, compared with the given sample size, some of the nonzero parameters may be moderately small, thereby fooling the consistent model selection procedure into choosing an incorrect model which then translates into bad risk behavior of the PMSE. Mathematically speaking, the problem is that the asymptotic loss-efficiency results discussed above are only pointwise results and do not hold uniformly with respect to the parameter θ: Kabaila (2002) showed, in a simple setting where (21) holds, that for given sample size n, there exists a parameter $\theta = \theta_n$ with infinitely many nonzero coordinates such that

$$\frac{L_n(\theta, \widehat{\theta}(\widehat{M}_{AIC}))}{L_n(\theta, \widehat{\theta}(\widehat{M}_{BIC}))} \geq 1,$$

and such that the ratio on the left-hand side in the above expression is greater than 2 with probability larger than 0.13. This shows that the results of Shibata (1981b), which entail that

$$\limsup_{n\to\infty} L_n(\theta, \widehat{\theta}(\widehat{M}_{AIC}))/L_n(\theta, \widehat{\theta}(\widehat{M}_{BIC})) \leq 1$$

for every *fixed* θ as $n \to \infty$, do not hold uniformly in θ (as for fixed n there exists a parameter θ for which the situation is reversed). Similarly, Shao's (1997) asymptotic loss-efficiency results for consistent procedures mentioned above are only pointwise asymptotic results and thus similarly problematic. Compare this with the discussion of the "oracle" phenomenon in Section 2.2 showing that consistent model selection procedures have bad maximal risk properties when selecting from a finite family of finite-dimensional models.

Given the dichotomy arising from the *pointwise* asymptotic loss-efficiency results, attempts have been made to devise "adaptive" model selection procedures that work well in both scenarios, i.e., procedures that combine the beneficial properties of both consistent and conservative procedures but avoid their detrimental properties. While this can be achieved in a *pointwise* asymptotic framework (Yang (2007); Ing (2007)), it is not surprising—given the preceding discussion—that it is impossible to achieve this goal in a uniform sense: No model selection procedure can simultaneously be consistent (like the BIC) and minimax-rate adaptive (like the AIC) as shown in Yang (2005). This is related to the fact that consistent model selection procedures lead to PMSEs that have maximal risk that diverges to infinity as sample size in-

creases, even for finite-dimensional models; see the discussion regarding (19) in Section 2.2.

The discussion so far again demonstrates that *pointwise* large-sample limit analyses of model selection procedures can paint a picture that is misleading in the sense that it need not have much resemblance to the situation in finite samples of *any* size. We now turn to two recent lines of research that analyze model selection procedures from a different perspective. Both of these lines of research rely on a combination of finite-sample results and asymptotic results that hold uniformly over (certain regions of) the parameter space, instead of pointwise asymptotic analyses.

In recent years, finite-sample risk bounds for PMSEs have been developed in considerable generality; see Barron and Cover (1991), Barron et al. (1999), and the references given in that paper. The following results are adapted from Birgé and Massart (2001) to our setting. Assume that the error variance σ^2 is known. For a (finite or countable) collection $\mathcal{M}_n$ of candidate models, consider the model selector $\widehat{M}$ that minimizes the following C_p-like criterion over $M \in \mathcal{M}_n$:

$$\mathsf{crit}(M) = \mathrm{RSS}(M) \ + \ \kappa(1 + \sqrt{2l_M})^2 \sigma^2 k_M.$$

This criterion depends on the user-specified constants $\kappa > 1$ and $l_M \geq 0$ that are chosen so that

$$\sum_{\substack{M \in \mathcal{M}_n, \\ k_M > 0}} e^{-l_M k_M} \leq \Sigma < \infty. \tag{22}$$

[For $\kappa(1 + \sqrt{2l_M})^2 = 2$ this criterion coincides with (6) up to an irrelevant additive constant.] Then the resulting PMSE $\widehat{\theta}(\widehat{M})$ has a risk $R_n(\theta, \widehat{\theta}(\widehat{M})) = \mathbb{E}_{n,\theta}[L_n(\theta, \widehat{\theta}(\widehat{M}))]$ satisfying

$$\begin{aligned} &R_n(\theta, \widehat{\theta}(\widehat{M})) \\ &\leq C(\kappa) n^{-1} \left[\inf_{M \in \mathcal{M}_n} \left(||(I - P_M) X\theta||^2 + \sigma^2 k_M (1 + l_M) \right) + \sigma^2 \Sigma \right], \end{aligned} \tag{23}$$

for $\theta \in l_2$, where the constant $C(\kappa)$ is given by $C(\kappa) = 12\kappa(\kappa+1)^3/(\kappa-1)^3$; cf. Theorem 2 and (3.12) in Birgé and Massart (2001), and observe that $(1+\sqrt{2x})^2 \leq 3(1+x)$ for $x \geq 0$. We stress that the upper bound in (23) holds under no additional assumptions on the unknown parameter θ other than $\theta \in l_2$. The upper bound in (23) equals $C(\kappa)$ times the sum of two terms. The first one is the infimum over all candidate models of a "penalized" version of the bias of the model M, i.e., $||(I - P_M)X\theta||^2/n$, where the "penalty" is given by $\sigma^2 k_M(1 + l_M)/n$. It should be noted that $R_n(\theta, \widehat{\theta}(M))$, i.e., the risk when fitting model M, is given by $R_n(\theta, \widehat{\theta}(M)) = ||(I - P_M)X\theta||^2/n + \sigma^2 k_M/n$. If, in addition, the constants l_M can be chosen to be bounded, i.e., $l_M \leq L$ for each $M \in \mathcal{M}_n$, while still satisfying (22), it follows from (23) that

$$R_n(\theta, \widehat{\theta}(\widehat{M})) \leq (1+L)C(\kappa)\left[\inf_{M\in\mathcal{M}_n} R_n(\theta, \widehat{\theta}(M)) + \frac{\sigma^2}{n}\Sigma\right]. \qquad (24)$$

Provided that $l_M \leq L$ for each $M \in \mathcal{M}_n$, we hence see that the risk of the PMSE $\widehat{\theta}(\widehat{M})$ is bounded by a constant multiple of the risk of the minimal-risk candidate model plus a constant. Suppose that one of the finite-dimensional candidate models, say, M_0, is a correct model for θ, and that M_0 is the smallest model with that property. Then $\inf_{M\in\mathcal{M}_n} R_n(\theta, \widehat{\theta}(M))$ is not larger than $\sigma^2 k_{M_0}/n$; in that case, the infimum in (24) is of the same order as $\sigma^2\Sigma/n$. Conversely, suppose that θ is infinite-dimensional. Then $\inf_{M\in\mathcal{M}_n} R_n(\theta, \widehat{\theta}(M))$ typically goes to zero slower than $1/n$, with the effect that $\inf_{M\in\mathcal{M}_n} R_n(\theta, \widehat{\theta}(M))$ is now the dominating factor in the upper bound in (24). Birgé and Massart (2001) argued that, without additional assumptions on the true parameter θ, the finite-sample upper bound in (24) is qualitatively the best possible (cf. the discussion leading up to (2.9) in that paper); see also Sects. 3.3.1 and 3.3.2 of that paper for a discussion of the choice of the constants κ and l_M in relation to the family of candidate models $\mathcal{M}_n$. It can furthermore be shown that the maximal risk of $\widehat{\theta}(\widehat{M})$ over certain regions Θ in the parameter space is not larger than a constant times the minimax risk over Θ, i.e.,

$$\sup_{\theta\in\Theta} R_n(\theta, \widehat{\theta}(\widehat{M})) \leq C(\Theta, \kappa, L) \inf_{\bar{\theta}} \sup_{\theta\in\Theta} R_n(\theta, \bar{\theta}), \qquad (25)$$

where the infimum is taken over all estimators $\bar{\theta}$, and where the constant $C(\Theta, \kappa, L)$ depends on the quantities indicated but not on sample size. Results of that kind hold, for example, in the case where the candidate models are the nested models $M(p)$ and the parameter set $\Theta \subseteq l_2$ is, after an appropriate reparameterization, a Sobolev or a Besov body (cf. Section 6 of Birgé and Massart (2001) or Section 5 of Barron et al. (1999)). We also note that the results of Barron et al. (1999) are more general and cover the Gaussian regression model discussed here as a special case; similar results continue to hold for other problems, including maximum likelihood density estimation, minimum $\mathbb{L}_1$ regression, and general projection estimators. For further results in that direction, see Barron (1991, 1998), Yang and Barron (1998), and Yang (1999). Furthermore, Birgé (2006) derived results similar to (23)–(25) for model selection based on preliminary tests. He found that the resulting PMSEs sometimes perform favorably compared with the estimators based on penalized maximum likelihood or penalized least squares considered above, but that the implementation of the preliminary tests can be difficult.

Risk bounds like (23)-(25) are sometimes called "oracle inequalities" in the literature (although there is no precise definition of this term). Informally, these bounds state that the risk of the PMSE is "not too far away" from the "risk target," i.e., from the risk corresponding to the model or to the estimator that an all-seeing oracle would choose.

Beran (1996) also studied the loss of PMSEs, but had a different focus. Instead of concentrating on oracle inequalities for the risk, Beran studied the problem of estimating the risk or loss of PMSEs; see also Kneip (1994), Beran and Dümbgen (1998), and Beran (2000). For the sake of simplicity, consider as the collection of candidate models at sample size n the set of all nested models of order up to n, i.e., $\mathcal{M}_n = \{M(p) : 0 \le p \le n\}$, assume again that the variance σ^2 is known, and let $\widehat{M}_{C_p}$ denote a (measurable) minimizer of Mallows's C_p objective function

$$\mathsf{MC}_n(M) = \mathrm{RSS}(M) + 2k_M\sigma^2 - n\sigma^2$$

over the set of candidate models. It then follows from Theorem 2.1 and Example 3 of Beran and Dümbgen (1998) that

$$\mathbb{E}_{n,\theta}\left|L_n(\theta,\widehat{\theta}(\widehat{M}_{C_p})) - \inf_{M\in\mathcal{M}_n} L_n(\theta,\widehat{\theta}(M))\right| \le \frac{C}{\sqrt{n}}\left(\sigma^2 + \sigma\sqrt{\theta'\frac{X'X}{n}\theta}\right) \tag{26}$$

and

$$\mathbb{E}_{n,\theta}\left|L_n(\theta,\widehat{\theta}(\widehat{M}_{C_p})) - n^{-1}\mathsf{MC}_n(\widehat{M}_{C_p})\right| \le \frac{C}{\sqrt{n}}\left(\sigma^2 + \sigma\sqrt{\theta'\frac{X'X}{n}\theta}\right), \tag{27}$$

where C is a constant independent of n, X, θ, and σ^2. The relation (26) is similar to (24) in that it relates the selected model to the best model. [Of course, the results in (24) and (26) are qualitatively different, but we shall not discuss the differences here. Beran and Dümbgen (1998) also provided a bound similar to (26) for the risk instead of the loss; moreover, they showed that an extension of $\widehat{\theta}(\widehat{M}_{C_p})$, which is based on smooth shrinkage, is asymptotically minimax over certain regions in parameter space.] The result in (27) differs from those discussed so far in the sense that it shows that the loss of the model selected by Mallows's C_p, i.e., $L_n(\theta,\widehat{\theta}(\widehat{M}_{C_p}))$, which is unknown in practice, can actually be estimated by n^{-1} times the value of the C_p objective function $\mathsf{MC}_n(\widehat{M}_{C_p})$, provided only that $\sqrt{n}$ is large in relation to $\sigma^2 + \sigma(\theta'X'X\theta/n)^{1/2}$. [Note that the upper bounds in (26) and (27), although unknown, can be estimated, because $\theta'X'X\theta/n$ can be estimated. The error variance σ^2 is assumed to be known here; in practice, σ^2 can often be estimated with reasonable accuracy.] A similar statement also holds for the risk corresponding to the model selected by C_p. The ability to actually estimate the risk or loss of the PMSE is important, because it allows for inference after model selection, like, e.g., the construction of asymptotically valid confidence balls. See Beran (1996, 2000) and Beran and Dümbgen (1998) for results in that direction. We also note that these papers allow for more general classes of candidate models and also for more general estimators, that is, for estimators based on smooth shrinkage.

So far, we have considered the fixed-design setting, i.e., the regressors were assumed to be nonrandom. The case of random design was studied by Baraud (2002), Wegkamp (2003), and Birgé (2004), based on the results of Barron et al. (1999). Leeb (2006b) gave results similar to (26) and (27) in the case of random design, where the loss is defined as squared-error loss for out-of-sample prediction, when a variant of the generalized cross-validation criterion (or the criterion (12)) is used for model selection. Predictive inference after model selection is studied in Leeb (2007).

4 Related Procedures Based on Shrinkage and Model Averaging

Classical shrinkage-type estimators include the James–Stein estimator and related methods (cf. James and Stein (1961); Strawderman and Cohen (1971)), or the ridge estimator (cf. Hoerl and Kennard (1970)). In recent years, there has been renewed interest in shrinkage-type estimators; examples include the bridge estimator of Frank and Friedman (1993), the nonnegative garrote of Breiman (1995), the lasso of Tibshirani (1996), the lasso-type estimators analyzed by Knight and Fu (2000), the smoothly clipped absolute deviation (SCAD) estimators proposed by Fan and Li (2001), or the adaptive lasso of Zou (2006). Many of these estimators just mentioned are instances of penalized maximum likelihood or least squares estimators.

Model averaging estimators—instead of selecting one candidate model and the corresponding estimator—form a weighted sum of the estimators corresponding to each of the candidate models where the weights typically are allowed to depend on the data. Model averaging estimators occur naturally in a Bayesian framework, where each model is weighted by its posterior probability. Good entry points into the considerable amount of literature on Bayesian model averaging are Hoeting et al. (1999) or Brown et al. (2002); see also the references given in these papers. Of course, model averaging methods have also been analyzed from a frequentist perspective; see, e.g., Buckland et al. (1997), Magnus (2002), Juditsky and Nemirovski (2000), Yang (2001, 2003), Hjort and Claeskens (2003), Danilov and Magnus (2004), Leung and Barron (2006), as well as Bunea et al. (2007).

Both shrinkage-type estimators and model averaging estimators can be regarded as extensions of PMSEs: Clearly, PMSEs can be viewed as special cases of shrinkage-type estimators, in the sense that PMSEs restrict (shrink) certain individual components of the parameter vector to zero. PMSEs can also be viewed as a particular case of model averaging estimators, where the weights are such that the selected model gets weight 1 and the other models get weight 0; cf. (2). This suggests that a number of phenomena that one can observe for PMSEs have counterparts in the larger class of shrinkage estimators or the class of estimators based on model averaging: Certain

shrinkage estimators like the SCAD of Fan and Li (2001) or the adaptive lasso of Zou (2006) can have a "sparsity property" in the sense that zero components of the true parameter are estimated as exactly 0 with probability approaching 1 as sample size increases (provided that the estimator's tuning parameter is chosen appropriately); consistent PMSEs have the same property. It is therefore not surprising that shrinkage estimators that have this "sparsity property" perform unfavorably in terms of worst-case risk in large samples (cf. the discussion in Section 2.2). In particular, the worst-case risk of shrinkage estimators that have the sparsity property increases to infinity with sample size; cf. Leeb and Pötscher (2008a). Also, the phenomena discussed in Section 2.3, namely, that the distribution is typically highly non-normal and that the cumulative distribution function of PMSEs cannot be estimated with reasonable accuracy, also occur with shrinkage estimators or model averaging estimators; cf. Leeb and Pötscher (2006a), Pötscher (2006), Pötscher and Leeb (2007), and Pötscher and Schneider (2007).

5 Further Reading

Apart from the expository articles already mentioned (Hocking (1976); Thompson (1978a, b); Amemiya (1980); Giles and Giles (1993); Hansen and Yu (2001); Rao and Wu (2001); Leeb and Pötscher (2005)), the article by DeGooijer et al. (1985) provides a survey of model selection in time series analysis; see also Chapter 5 of Hannan and Deistler (1988). The Bayesian approach to model selection is discussed in Hoeting et al. (1999) and Berger and Pericchi (2001).

The books by Judge and Bock (1978), Linhart and Zucchini (1986), Choi (1992), McQuarrie and Tsai (1998), Burnham and Anderson (2002), Miller (2002), Saleh (2006), and Konishi and Kitagawa (2008) deal with various aspects of model selection.

There is also a considerable body of literature on model selection and related methods in the areas of machine learning and empirical risk minimization, mainly focusing on classification and pattern recognition problems; see, e.g., Boucheron et al. (2005) and Cesa-Bianchi and Lugosi (2006).

We finally mention a development that circles around the idea of automated discovery and automated modeling; for an introduction see Phillips (2005) and references therein.

References

Ahmed, S. E. and Basu, A. K. (2000): Least squares, preliminary test and Stein-type estimation in general vector AR(p) models. *Statistica Neerlandica* **54**, 47–66.

Akaike, H. (1969): Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

Akaike, H. (1970): Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. *In: B.N. Petrov and F. Csaki (Eds.): Second International Symposium on Information Theory*. Akadémiai Kiadó, Budapest.

Allen, D. M. (1971): Mean square error of prediction as a criterion for selecting variables. *Technometrics* **13**, 469–475.

Allen, D. M. (1974): The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125–127.

Amemiya, T. (1980): Selection of regressors. *International Economic Review* **21**, 331–354.

An, H. Z. and Chen, Z. G. (1986): The identification of ARMA processes. *Journal of Applied Probability* **23A**, 75–87.

An, H. Z. and Gu, L. (1985): On the selection of regression variables. *Acta Mathematicae Applicatae Sinica* **2**, 27–36.

Anderson, T. W. (1962): The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics* **33**, 255–265.

Anderson, T. W. (1963): Determination of the order of dependence in normally distributed time series. *In: Rosenblatt, M. (Ed.): Time Series Analysis*, 425–446. Wiley, New York.

Bancroft, T. A. and Han, C. P. (1977): Inference based on conditional specification: A note and a bibliography. *International Statistical Review* **45**, 117–127.

Baraud, Y. (2002): Model selection for regression on a random design. *ESAIM Probability and Statistics* **6**, 127–146.

Barron, A. R. (1991): Complexity regularization with application to artificial neural networks. *In: Nonparametric functional estimation and related topics (Spetses, 1990), NATO Advanced Study Institute Series C Mathematical and Physical Sciences* **335**, 561–576. Kluwer, Dordrecht.

Barron, A. R. (1999): Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. *In: Bayesian Statistics, (Alcoceber, 1998)* **6**, 27–52. Oxford University Press, Oxford.

Barron, A. R. and Cover, T. M. (1991): Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37**, 1034-1054. (Corrections: *IEEE Transactions on Information Theory* **37**, 1738.)

Barron, A. R., Birgé, L. and Massart, P. (1999): Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113**, 301–413.

Bauer, P., Pötscher, B. M. and Hackl, P. (1988): Model selection by multiple test procedures. *Statistics* **19**, 39–44.

Beran, R. (1996): Confidence sets centered at C_p-estimators. *Annals of the Institute of Statistical Mathematics* **48**, 1–15.

Beran, R. (2000): REACT scatterplot smoothers: Superefficiency through basis economy. *Journal of the American Statistical Association* **95**, 155–171.

Beran, R. and Dümbgen, L. (1998): Modulation of estimators and confidence sets. *Annals of Statistics* **26**, 1826–1856.

Berger, J. O. and Pericchi, L. R. (2001): Objective Bayesian methods for model selection: Introduction and comparison. *In: Lahiri, P. (Ed.): Model Selection. IMS Lecture Notes Monograph Series* **38**, 135–193.

Bhansali, R. J. (1999): Parameter estimation and model selection for multistep prediction of a time series: A review. *In: Ghosh, S. (Ed.): Asymptotics, Nonparametrics and Time Series—A Tribute to Madan Lal Puri*, 201–225. Dekker, New York.

Bhansali, R. J. and Downham, D. Y. (1977): Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547–551.

Birgé, L. (2004): Model selection for Gaussian regression with random design. *Bernoulli* **10**, 1039–1051.

Birgé, L. (2006): Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'Institute Henri Poincaré* **42**, 273–325.

Birgé, L. and Massart, P. (2001): Gaussian model selection. *Journal of the European Mathematical Society* **3**, 203–268.

Breiman, L. (1995): Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Breiman, L. and Freedman, D. (1983): How many variables should be entered in a regression equation? *Journal of the American Statistical Association* **78**, 131–136.

Brook, R. J. (1976): On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Association* **71**, 126-131. (Correction: *Journal of the American Statistical Association* **71**, 1010.)

Brown, P. J., Vannucci, M. and Fearn, T. (2002): Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society Series B* **64**, 519–536.

Boucheron, S., Bousquet, O. and Lugosi, G. (2005): Theory of classification: A survey of some recent advances. *ESAIM Probability and Statistics* **9**, 323–375.

Buckland S. T., Burnham, K. P. and Augustin, N. H. (1997): Model selection: An integral part of inference. *Biometrics* **53**, 603–618.

Bunea, F. (2004): Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics* **32**, 898–927.

Bunea, F., Tsybakov, A. and Wegkamp, M. H. (2007): Aggregation for Gaussian regression. *Annals of Statistics* **35**, 1674–1697.

Bunea, F., Wegkamp, M. H. and Auguste, A. (2006): Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* **136**, 4349–4364.

Burnham, K. P. and Anderson, D. R. (2002): *Model Selection and Multimodal Inference* (2nd edition). Springer, New York.

Cesa-Bianchi, N. and Lugosi, G. (2006): *Prediction, Learning, and Games.* Cambridge University Press, Cambridge.

Chan, N.-H. (2008): Time series with roots on or near the unit circle. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*, 694–707. Springer, New York.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998): Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20**, 33–61.

Chen, Z. G. and Ni, J. Y. (1989): Subset regression time series and its modeling procedures. *Journal of Multivariate Analysis* **31**, 266–288.

Choi, B. (1992): *ARMA Model Identification.* Springer, New York.

Claeskens, G. and Hjort, N. L. (2003): The focused information criterion. *Journal of the American Statistical Association* **98**, 900–916.

Craven, P. and Wahba, G. (1979): Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

Danilov, D. and Magnus, J. R. (2004): On the harm that ignoring pretesting can cause. *Journal of Econometrics* **122**, 27–46.

Davisson, L. D. (1965): The prediction error of stationary Gaussian time series of unknown covariance. *IEEE Transactions on Information Theory* **11**, 527–532.

DeGooijer, J. G., Bovas, A., Gould, A. and Robinson, L. (1985): Methods for determining the order of an autoregressive-moving average process: A survey. *International Statistical Review* **53**, 301–329.

Dijkstra, T. K. and Veldkamp, J. H. (1988): Data-driven selection of regressors and the bootstrap. *In: Dijkstra, T. K. (Ed.): Lecture Notes in Economics and Mathematical Systems* **307**, 17–38, Springer, New York.

Draper, N. R. and Smith, H. (1981): *Applied Regression Analysis* (2nd edition). Wiley, New York.

Droge, B. (1993): On finite-sample properties of adaptive least squares regression estimates. *Statistics* **24**, 181–203.

Droge, B. and Georg, T. (1995): On selecting the smoothing parameter of least squares regression estimates using the minimax regret approach. *Statistics and Decisions* **13**, 1–20.

Dufour, J. M., Pelletier, D. and Renault, E. (2006): Short run and long run causality in time series: Inference. *Journal of Econometrics* **132**, 337–362.

Fan, J. and Li, R. (2001): Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Findley, D. F. (1985): On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *Journal of Time Series Analysis* **6**, 229–252.

Findley, D. F. (1991): Model selection for multistep-ahead forecasting. *American Statistical Association Proceedings of the Business and Economic Statistics Section* 243–247.

Findley, D. F. and Wei, C. Z. (2002): AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *Journal of Multivariate Analysis* **83**, 415–450.

Foster, D. P. and George, E. I. (1994): The risk inflation criterion for multiple regression. *Annals of Statistics* **22**, 1947–1975.

Frank, I. E. and Friedman, J. H. (1993): A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.

Francq, C., Roussignol, M. and Zakoïan, J. M. (2001): Conditional heteroskedasticity driven by hidden Markov chains. *Journal of Time Series Analysis* **22**, 197–220.

George, E. I. and Foster, D. P. (2000): Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

Geweke, J. and Meese, R. (1981): Estimating regression models of finite but unknown order. *International Economic Review* **22**, 55–70.

Giles, J. A. and Giles, D. E. A. (1993): Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys* **7**, 145–197.

Guyon, X. and Yao, J. (1999): On the underfitting and overfitting sets of models chosen by order selection criteria. *Journal of Multivariate Analysis* **70**, 221–249.

Hall, A. R. and Peixe, F. P. M. (2003): A consistent method for the selection of relevant instruments. *Econometric Reviews* **22**, 269–287.

Hannan, E. J. (1980): The estimation of the order of an ARMA process. *Annals of Statistics* **8**, 1071–1081.

Hannan, E. J. (1981): Estimating the dimension of a linear system. *Journal of Multivariate Analysis* **11**, 459–473.

Hannan, E. J. and Deistler, M. (1988): *The Statistical Theory of Linear Systems*. Wiley, New York.

Hannan, E. J. and Quinn, B. G. (1979): The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B* **41**, 190–195.

Hansen, M. H. and Yu, B. (2001): Model selection and the principle of minimum description length. *Journal of the American Statistical Association* **96**, 746–774.

Haughton, D. (1991): Consistency of a class of information criteria for model selection in nonlinear regression. *Communications in Statistics. Theory and Methods* **20**, 1619–1629.

Hemerly, E. M. and Davis, M.H.A. (1989): Strong consistency of the PLS criterion for order determination of autoregressive processes. *Annals of Statistics* **17**, 941–946.

Hidalgo, J. (2002): Consistent order selection with strongly dependent data and its application to efficient estimation. *Journal of Econometrics* **110**, 213–239.

Hjort, N. L. and Claeskens, G. (2003): Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Hocking, R. R. (1976): The analysis and selection of variables in linear regression. *Biometrics* **32**, 1–49.

Hoerl, A. E. and Kennard, R. W. (1970): Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Hoeting, J. A., Madigan, D., Raftery, A. and Volinsky, C. T. (1999): Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382–401. (Corrections: *Statistical Science* **15**, 193–195.)

Hosoya, Y. (1984): Information criteria and tests for time series models. *In: Anderson, O. D. (Ed.): Time Series Analysis: Theory and Practice* **5**, 39–52. North-Holland, Amsterdam.

Hosoya, Y. (1986): A simultaneous test in the presence of nested alternative hypotheses. *Journal of Applied Probability* **23A**, 187–200.

Hurvich, M. M. and Tsai, C. L. (1989): Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ing, C. K. (2004): Selecting optimal multistep predictors for autoregressive processes of unknown order. *Annals of Statistics* **32**, 693–722.

Ing, C. K. (2007): Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics* **35**, 1238–1277.

Ing, C. K. and Wei, C. Z. (2005): Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics* **33**, 2423–2474.

Ing, C. K. and Wei, C. Z. (2006): A maximal moment inequality for long range dependent time series with applications to estimation and model selection. *Statistica Sinica* **16**, 721–740.

James, W. and Stein, C. (1961): Estimation with quadratic loss. *In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 361–379. California University Press, Berkeley.

Judge, G. G. and Bock, M. E. (1978): *The Statistical Implications of Pre-test and Stein-Rule Estimators in Econometrics*. North-Holland, Amsterdam.

Juditsky, A. and Nemirovski, A. (2000): Functional aggregation for nonparametric regression. *Annals of Statistics* **28**, 681–712.

Kabaila, P. (1995): The effect of model selection on confidence regions and prediction regions. *Econometric Theory* **11**, 537–549.

Kabaila, P. (1996): The evaluation of model selection criteria: Pointwise limits in the parameter space. *In: Dowe, D. L., Korb, K. B. and Oliver, J. J. (Eds.): Information, Statistics and Induction in Science*, 114–118. World Scientific, Singapore.

Kabaila, P. (1998): Valid confidence intervals in regression after variable selection. *Econometric Theory* **14**, 463–482.

Kabaila, P. (2002): On variable selection in linear regression. *Econometric Theory* **18**, 913–925.

Kabaila, P. and Leeb, H. (2006): On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* **101**, 619–629.

Kempthorne, P. J. (1984): Admissible variable-selection procedures when fitting regression models by least squares for prediction. *Biometrika* **71**, 593–597.

Kennedy, W. J. and Bancroft, T. A. (1971): Model building for prediction in regression based upon repeated significance tests. *Annals of Mathematical Statistics* **42**, 1273–1284.

Keribin, C. and Haughton, D. (2003): Asymptotic probabilities of over-estimating and under-estimating the order of a model in general regular families. *Communications in Statistics. Theory and Methods* **32**, 1373–1404.

Kneip, A. (1994): Ordered linear smoothers. *Annals of Statistics* **22**, 835–866.

Knight, K. (1989): Consistency of Akaike's information criterion for infinite variance autoregressive processes. *Annals of Statistics* **17**, 824–840.

Knight, K. and Fu, W. (2000): Asymptotics for lasso-type estimators. *Annals of Statistics* **28**, 1356–1378.

Kohn, R. (1983): Consistent estimation of minimal subset dimension. *Econometrica* **51**, 367–376.

Konishi, S. and Kitagawa, G. (1996): Generalized information criteria in model selection. *Biometrika* **83**, 875–890.

Konishi, S. and Kitagawa, G. (2008): *Information Criteria and Statistical Modeling*. Springer, New York.

Leeb, H. (2005): The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximations. *Journal of Statistical Planning and Inference* **134**, 64–89.

Leeb, H. (2006a): The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations. *In: Rojo, J. (Ed.): IMS Lecture Notes Monograph Series* **49**, 291–311. Institute of Mathematical Statistics, Beachwood.

Leeb, H. (2006b): Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, forthcoming.

Leeb, H. (2007): Conditional predictive inference post model selection. *Manuscript, Department of Statistics, Yale University*.

Leeb, H. and Pötscher, B. M. (2003): The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* **19**, 100–142.

Leeb, H. and Pötscher, B. M. (2005): Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.

Leeb, H. and Pötscher, B. M. (2006a): Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* **22**, 69–97. (Corrigendum. *Econometric Theory* **24**, 581–583.)

Leeb, H. and Pötscher, B. M. (2006b): Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* **34**, 2554–2591.

Leeb, H. and Pötscher, B. M. (2008a): Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142**, 201–211.

Leeb, H. and Pötscher, B. M. (2008b): Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24**, 338–376.

Leung, G. and Barron, A. R. (2006): Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52**, 3396–3410.

Li, K. C. (1987): Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* **15**, 958–975.

Linhart, H. and Zucchini, W. (1986): *Model Selection*. Springer, New York.

Lütkepohl, H. (1990): Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics* **72**, 116–125.

Magnus, J. R. (1999): The traditional pretest estimator. *Teoriya Veroyatnostei i Ee Primeneniya* **44**, 401–418; translation in *Theory of Probability and Its Applications* **44**, (2000), 293–308.

Magnus, J. R. (2002): Estimation of the mean of a univariate normal distribution with known variance. *The Econometrics Journal* **5**, 225–236.

Mallows, C. L. (1965): Some approaches to regression problems. *Unpublished manuscript*.

Mallows, C. L. (1967): Choosing a subset regression. *Bell Telephone Laboratories unpublished report*.

Mallows, C. L. (1973): Some comments on C_p. *Technometrics* **15**, 661–675.

Mallows, C. L. (1995): More comments on C_p. *Technometrics* **37**, 362–372.

McKay, R. J. (1977): Variable selection in multivariate regression: An application of simultaneous test procedures. *Journal of the Royal Statistical Society Series B* **39**, 371–380.

McQuarrie, A. D. R. and Tsai, C. L. (1998): *Regression and time series model selection*. World Scientific, River Edge.

Miller, A. (2002): *Subset Selection in Regression* (2nd edition). Chapman and Hall, Boca Raton.

Nishii, R. (1984): Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* **12**, 758–765.

Nishii, R. (1988): Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis* **27**, 392–403.

Paulsen, J. (1984): Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* **5**, 115–127.

Phillips, P. C. B. (2005): Automated discovery in econometrics. *Econometric Theory* **21**, 3–20.

Polyak, B. T. and Tsybakov, A. B. (1990): Asymptotic optimality of the C_p -test for the orthogonal series estimation of regression. *Theory of Probability and Its Applications* **35**, 293–306.

Pötscher, B. M. (1983): Order estimation in ARMA-models by Lagrangian multiplier tests. *Annals of Statistics* **11**, 872–885.

Pötscher, B. M. (1985): The behaviour of the Lagrangian multiplier test in testing the orders of an ARMA-model. *Metrika* **32**, 129–150.

Pötscher, B. M. (1989): Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Annals of Statistics* **17**, 1257–1274.

Pötscher, B. M. (1990): Estimation of autoregressive moving average order given an infinite number of models and approximation of spectral densities. *Journal of Time Series Analysis* **11**, 165–179.

Pötscher, B. M. (1991): Effects of model selection on inference. *Econometric Theory* **7**, 163–185.

Pötscher, B. M. (1995): Comment on 'The effect of model selection on confidence regions and prediction regions' by P. Kabaila. *Econometric Theory* **11**, 550–559.

Pötscher, B. M. (2006): The distribution of model averaging estimators and an impossibility result regarding its estimation. *In: Ho, H.-C., Ing, C.-K. and Lai, T.-L. (Eds.): Time Series and Related Topics: In Memory of Ching-Zong Wei. IMS Lecture Notes and Monograph Series* **52**, 113–129. Institute of Mathematical Statistics, Beachwood.

Pötscher, B. M. (2007): Confidence sets based on sparse estimators are necessarily large. *Working paper, Department of Statistics, University of Vienna.* arXiv:0711.1036.

Pötscher, B. M. and Leeb, H. (2007): On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Working paper, Department of Statistics, University of Vienna.* arXiv:0711.0660.

Pötscher, B. M. and Novak, A. J. (1998): The distribution of estimators after model selection: Large and small sample results. *Journal of Statistical Computation and Simulation* **60**, 19–56.

Pötscher, B. M. and Schneider, U. (2007): On the distribution of the adaptive LASSO estimator. *Working paper, Department of Statistics, University of Vienna.* arXiv:0801.4627.

Pötscher, B. M. and Srinivasan, S. (1994): A comparison of order estimation procedures for ARMA models. *Statistica Sinica* **4**, 29–50.

Quinn, B. G. (1980): Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society Series B* **42**, 182–185.

Quinn, B. G. (1988): A note on AIC order determination for multivariate autoregressions. *Journal of Time Series Analysis* **9**, 241–245.

Rao, C. R. and Wu, Y. (1989): A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369–374.

Rao, C. R. and Wu, Y. (2001): On model selection. *In: Lahiri, P. (Ed.): Model Selection. IMS Lecture Notes Monograph Series* **38**, 1–57. Institute of Mathematical Statistics, Beachwood.

Rao, C. R. and Wu, Y. (2005): Linear model selection by cross-validation. *Journal of Statistical Planning and Inference* **128**, 231–240.

Reschenhofer, E. (1999): Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econometric Theory* **15**, 377–387.

Rissanen, J. (1978): Modeling by shortest data description. *Automatica* **14**, 465–471.

Rissanen, J. (1983): A universal prior for integers and estimation by minimum description length. *Annals of Statistics* **11**, 416–431.

Rissanen, J. (1986a): Stochastic complexity and modeling. *Annals of Statistics* **14**, 1080–1100.

Rissanen, J. (1986b): A predictive least squares principle. *IMA Journal of Mathematical Control and Information* **3**, 211–222.

Rissanen, J. (1987): Stochastic complexity (with discussion). *Journal of the Royal Statistical Society Series B* **49**, 223–265.

Rissanen, J. (1989): *Stochastic Complexity and Statistical Inquiry*. World Scientific, Teaneck.

Sakai, H. (1981): Asymptotic distribution of the order selected by AIC in multivariate autoregressive model fitting. *International Journal of Control* **33**, 175–180.

Saleh, A. K. M. E. (2006): *Theory of Preliminary Test and Stein-Type Estimation with Applications*. Wiley, Hoboken.

Saleh, A. K. M. E. and Sen, P. K. (1983): Asymptotic properties of tests of hypothesis following a preliminary test. *Statistics and Decisions* **1**, 455–477.

Sawa, T. and Hiromatsu, T. (1973): Minimax regret significance points for a preliminary test in regression analysis. *Econometrica* **41**, 1093–1101.

Schwarz, G. (1978): Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sclove, S. L., Morris, C. and Radhakrishnan, R. (1972): Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* **43**, 1481–1490.

Sen, P. K (1979): Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* **7**, 1019–1033.

Sen, P. K and Saleh, A. K. M. E. (1987): On preliminary test and shrinkage M-estimation in linear models. *Annals of Statistics* **15**, 1580–1592.

Shao, J. (1993): Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Shao, J. (1997): An asymptotic theory for linear model selection (with discussion) *Statistica Sinica* **7**, 221–264.

Shao, J. (1998): Convergence rates of the generalized information criterion. *Journal of Nonparametric Statistics* **9**, 217–225.

Shen, X., Huang, H. C. and Ye, J. (2004): Inference after model selection. *Journal of the American Statistical Association* **99**, 751–762.

Shibata, R. (1976): Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–126.

Shibata, R. (1980): Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* **8**, 147–164.

Shibata, R. (1981a): An optimal autoregressive spectral estimate. *Annals of Statistics* **9**, 300–306.

Shibata, R. (1981b): An optimal selection of regression variables. *Biometrika* **68**, 45–54. (Correction: *Biometrika* **69**, 492).

Shibata, R. (1984): Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43–49.

Shibata, R. (1986a): Consistency of model selection and parameter estimation. *Journal of Applied Probability* **23A**, 127–141.

Shibata, R. (1986b): Selection of the number of regression variables; a minimax choice of generalized FPE. *Annals of the Institute of Statistical Mathematics* **38**, 459–474.

Shibata, R. (1989): Statistical aspects of model selection. *In: J. C. Willems (Ed.): From Data to Model*, 215–240. Springer, New York.

Shibata, R. (1997): Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica* **7**, 375–394.

Söderström, T. (1977): On model structure testing in system identification. *International Journal of Control* **26**, 1–18.

Stone, C. (1981): Admissible selection of an accurate and parsimonious normal linear regression model. *Annals of Statistics* **9**, 475–485.

Stone, C. (1982): Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Annals of the Institute of Statistical Mathematics* **34**, 123–133.

Stone, M. (1974): Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society Series B* **36**, 111–133.

Stone, M. (1977): An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B* **39**, 44–47.

Strawderman, W. E. (1971): Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics* **42**, 385–388.

Sugiura, N. (1978): Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics* **A7**, 13–26.

Takada, Y. (1982): Admissibility of some variable selection rules in linear regression model. *Journal of the Japanese Statistical Society* **12**, 45–49.

Takeuchi, K. (1976): Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* **153**, 12–18. (In Japanese.)

Teräsvirta, T. and Mellin, I. (1986): Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics* **13**, 159–171.

Theil, H. (1961): *Economic Forecasts and Policy* (2nd edition). North-Holland, Amsterdam.

Thompson, M. L. (1978a): Selection of variables in multiple regression: part I. A review and evaluation. *International Statistical Review* **46**, 1–19.

Thompson, M. L. (1978b): Selection of variables in multiple regression: part II. Chosen procedures, computations and examples. *International Statistical Review* **46**, 129–146.

Tibshirani, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Toro-Vizcarrondo, C. and Wallace, T. D. (1968): A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association* **63**, 558–572.

Toyoda, T. and Wallace, T. D. (1976): Optimal critical values for pre-testing in regression. *Econometrica* **44**, 365–375.

Tsay, R. S. (1984): Order selection in nonstationary autoregressive models. *Annals of Statistics* **12**, 1425–1433.

Venter, J. H. and Steele, S. J. (1992): Some contributions to selection and estimation in the normal linear model. *Annals of the Institute of Statistical Mathematics* **44**, 281–297.

Vuong, Q. H. (1989): Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333.

Wallace, T. D. (1972): Weaker criteria and tests for linear restrictions in regression. *Econometrica* **40**, 689–698.

Wei, C. Z. (1992): On predictive least squares principles. *Annals of Statistics* **20**, 1-42.

Wegkamp, M. (2003): Model selection in nonparametric regression. *Annals of Statistics* **31**, 252–273.

Yang, Y. (1999): Model selection for nonparametric regression. *Statistica Sinica* **9**, 475–499.

Yang, Y. (2001): Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 574–588.

Yang, Y. (2003): Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* **13**, 783–809.

Yang, Y. (2005): Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937–950.

Yang, Y. (2007): Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory* **23**, 1–36.

Yang, Y. and Barron, A. R. (1998): An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* **44**, 95–116.

Yang, Y. and Barron, A. R. (1999): Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* **27**, 1564–1599.

Zhang, P. (1992): Inference after variable selection in linear regression models. *Biometrika* **79**, 741–746.

Zhang, P. (1993a): Model selection via multifold cross validation. *Annals of Statistics* **21**, 299–313.

Zhang, P. (1993b): On the convergence rate of model selection criteria. *Communications in Statistics. Theory and Methods* **22**, 2765–2775.

Zheng, X. and Loh, W. Y. (1995): Consistent variable selection in linear models. *Journal of the American Statistical Association* **90**, 151–156.

Zheng, X. and Loh, W. Y. (1997): A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica* **7**, 311–325.

Zou, H. (2006): The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Nonparametric Modeling in Financial Time Series

Jürgen Franke, Jens-Peter Kreiss and Enno Mammen

Abstract In this chapter, we deal with nonparametric methods for discretely observed financial data. The main ideas of nonparametric kernel smoothing are explained in the rather simple situation of density estimation and regression. For financial data, a rather relevant topic is nonparametric estimation of a volatility function in a continuous-time model such as a homogeneous diffusion model. We review results on nonparametric estimation for discretely observed processes, sampled at high or at low frequency. We also discuss application of nonparametric methods to testing, especially model validation and goodness-of-fit testing. In risk measurement for financial time series, conditional quantiles play an important role and nonparametric methods have been successfully applied in this field too. At the end of the chapter we discuss Grenander's sieve methods and other more recent advanced nonparametric approaches.

1 Introduction

Flexible stochastic modeling is a relevant topic not only for financial time series, and nonparametric methods offer a unified approach for statistical inference in many fields of applications. Firstly developed for independent data,

Jürgen Franke
Department of Mathematics, Universität Kaiserslautern, Erwin–Schrödinger-Strasse, 67653 Kaiserslautern, Germany, e-mail: franke@mathematik.uni-kl.de

Jens-Peter Kreiss
Institut für Mathematische Stochastik, Technische Universität Braunschweig, Pockelsstrasse 14, 38106 Braunschweig, Germany, e-mail: j.kreiss@tu-bs.de

Enno Mammen
Abteilung Volkswirtschaftslehre, L7, 3–5, Universität Mannheim, 68131 Mannheim, Germany, e-mail: emammen@rumms.uni-mannheim.de

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_40,

nonparametric methods play a growing role in the area of dependent data as well as in time series analysis. In time series analysis, one early reference regarding nonparametric kernel estimation is Robinson (1983). He stated multivariate central limit theorems for estimators of finite-dimensional densities of a strictly stationary process. Furthermore, he studied the asymptotics for estimators of conditional densities and conditional expectations under strong mixing conditions. The literature on nonparametric financial time series analysis is not restricted to kernel smoothing and its generalizations (like local polynomial estimators, LPEs). Other approaches that have been considered in the literature include smoothing splines, orthogonal series estimators and curve estimators from the area of learning theory. In this chapter we mainly focus on kernel and local polynomial smoothing methods for time series.

Besides estimation of relevant statistical quantities, nonparametric methods offer a general approach for testing, especially for goodness-of-fit tests, model checking and validation. A widely implemented idea uses comparisons of completely nonparametric estimators with parametric or semiparametric model based counterparts. The null hypothesis of the parametric or semiparametric model is rejected if the two estimators differ too much. Also, in a rather informal way nonparametric methods may be viewed as a diagnostic tool for model building.

This chapter starts with a general description of smoothing methods for time series (Section 2). In Section 2.1, we explain general ideas in nonparametric density estimation. Section 2.2 then investigates the more complex nonparametric estimation of a conditional expectation.

So far, the underlying data are discretely observed time series. Of course for financial time series, the underlying modeling quite often is done in continuous time. However, also for a continuous-time model, typically the data are discretely observed. Section 2.3 treats nonparametric methods for discretely observed financial continuous-time models. In particular, we discuss nonparametric methods for diffusions. A main problem here is the estimation of the volatility function. Estimation is quite different for data that are observed with high or with low frequency, respectively.

In Section 3, we come back to the already mentioned application of nonparametric methods to the area of testing and model checking. Another relevant application of nonparametric methods is nonparametric quantile estimation. Conditional quantiles are an important tool for defining risk measures on the basis of financial time series. The famous value at risk (VaR), for example, is given directly by a quantile of the return distribution, but also the so-called expected shortfall as a coherent measure of risk uses conditional quantiles in its definition as expected excess loss above the VaR. We give an overview of this field in Section 4.

Section 5 is on advanced nonparametric modeling, where we mainly deal with dimension reduction techniques like additive models and functional principal components and with nonstationary models. The final section discusses

Grenander's sieve methods (Section 6) and concludes this chapter on nonparametric methods and their application to financial time series.

For further reviews on nonparametric methods for time series, including applications in finance and financial econometrics, we refer the interested reader to the following monographs and overview articles: Härdle and Linton (1994), Pagan and Ullah (1999), Fan and Yao (2003), Fan (2005), Gao (2007) and Zhao (2008a). Readers mainly interested in nonparametric and semiparametric generalizations of autoregressive (AR) conditional heteroscedasticity (ARCH) modeling are also referred to the chapter by Linton (2008) in this handbook.

2 Nonparametric Smoothing for Time Series

In this section, we deal with the main principles and results of kernel smoothing methods in the field of time series analysis. In order to keep the situation as simple as possible, but still relevant, we are going to start with the fundamental problem of density estimation. For this problem, the main ideas of kernel smoothing can be formulated without major technical difficulties. It is argued that under suitable assumptions regarding the dependence structure of the underlying time series data, we more or less obtain results comparable to the classical independent and identically distributed situation. Besides this, we briefly review some applications of density estimation in the area of finance.

2.1 Density estimation via kernel smoothing

Let us assume that we have observations $X_1, \ldots, X_n$ of a stationary and real-valued univariate time series at hand. We assume that the stationary distribution exists and that it possesses a continuous density p, say. It is intended to estimate this density p on the basis of our data. One standard density estimator is the so-called *Nadaraya–Watson* kernel density estimator, which has the following form

$$\widehat{p}(x) = \frac{1}{nh} \sum_{t=1}^{n} K\left(\frac{x - X_t}{h}\right) . \tag{1}$$

Here, x denotes a fixed value in $\mathbb{R}$, K is a probability density (called *kernel*) and $h > 0$ denotes a so-called *bandwidth*. The bandwidth is a crucial parameter in kernel smoothing and has to be chosen with great care. Typically, one chooses the kernel K such that it has a compact support. In many cases $K : [-1, +1] \to \mathbb{R}$; often-used kernels on $[-1, +1]$ are

$K(u) = 3/4\ (1-u)^2$ (*Epanechnikov kernel*) and $K(u) = 15/16\ (1-u^2)^2$, the so-called *biweight* or *quartic kernel*. Kernels with unbounded support are $K(u) = 1/\sqrt{2\pi}\exp(-u^2/2)$, the *standard normal kernel*, or the *logistic kernel* $K(u) = \exp(u)/(1+\exp(u))^2$.

For compactly supported kernel K, it holds that the summands of the Nadaraya–Watson kernel density smoother are nonvanishing if and only if an observation X_t falls in a small neighborhood of the fixed value x. If one ensures through an assumption regarding the dependence structure of the underlying data that observations falling in such an neighborhood are fairly separated in time, it is reasonable to expect that the asymptotics of the Nadaraya–Watson density estimator for time series data satisfying such a dependence assumption do not significantly differ from the asymptotics in the much simpler independent and identically distributed setting.

A classical assumption for a dependence structure is, of course, mixing. The notion of strongly mixing observations, which has been used in many applications, is defined as follows.

Definition 1 (Section 1.1 in Bosq (1996))
A process $(X_t : t \in \mathbb{Z}))$ is said to be α-*mixing* or *strongly mixing* if

$$\alpha_k := \sup_{t\in\mathbb{Z}} \left(\sup\left(|P(B\cap C) - P(B)P(C)|\right)\right) \ , \tag{2}$$

where the second "sup" is over all sets B and C, with $B \in \sigma(X_s, s \le t)$ and $C \in \sigma(X_s, s \ge t+k)$.
The "$\sup_{t\in\mathbb{Z}}$" may be omitted if (X_t) is stationary.

For observations that are strongly mixing and strictly stationary we have the following result for the asymptotic distribution of the Nadaraya–Watson kernel density estimator.

Theorem 1 (Theorem 2.3 in Bosq (1996))
Let $(X_t : t \in \mathbb{Z})$ denote a real-valued strictly stationary process. Assume that whenever $t_1 < t_2 < t_3 < t_4$ $(X_{t_1}, \ldots, X_{t_4})$ possesses a density, which is almost surely uniformly bounded (i.e., $\|\sup_{t_1,\ldots,t_4} p_{t_1,\ldots,t_4}\|_\infty$, that for the two-dimensional densities $p_{s,t}$ and the univariate stationary density p $\sup_{s<t}\|f_{s,t} - f\otimes f\|_\infty$ holds and that p is bounded and twice continuously differentiable with a bounded second derivative. Furthermore, if $\alpha_k = \mathcal{O}(k^{-\beta})$, where $\beta \ge 2$, and if $h_n = \frac{c}{\log\log n} n^{-1/5}, c > 0$, then for all integers m and all distinct $x_1, \ldots, x_m$ such that $f(x_i) > 0$

$$(nh_n)^{1/2}\left(\frac{\widehat{p}(x_i) - p(x_i)}{\left(\widehat{p}(x_i)\int K^2(u)du\right)^{1/2}}, 1 \le i \le m\right) \to \mathcal{N}(0,1)^m, \tag{3}$$

where $\mathcal{N}(0,1)^m$ denotes the m-dimensional standard normal distribution.

From Theorem 1 it is easily seen that $\widehat{p}(x)$ and $\widehat{p}(y)$ for $x \ne y$ are asymptotically independent. Heuristically, this can be explained as follows for kernels

K with compact support. For h small enough, the sets of observations—which enter into the computation of these two estimators—are disjoint. Furthermore, if one considers the time lag between two observations, one falling into the neighborhood of x and the other one falling into the neighborhood of y, one observes that this time lag typically is rather large. The assumptions regarding the decay rate of the strong mixing coefficients then imply asymptotic independence. Even more, the time lag between observations falling in one and the same neighborhood around x, say, is also large. By the same reasoning as above, this leads to the fact that the asymptotics of the kernel density estimator for strongly mixing time series observations are exactly the same as for independent and identically distributed observations. This holds as long as the mixing coefficients decay sufficiently fast. Hart (1996) coined for this fact the term "whitening by windowing."

The result of Theorem 1 can easily be extended to bandwidths h with other rates of convergence. If the bandwidth h is of order $n^{-1/5}$ or slower, an additional bias term of order h^2 shows up and is not asymptotically negligible as it is in Theorem 1 .

The problem of nonparametric density estimation has been dealt with extensively in the literature. An early reference is Collomb (1984), but there are lots of others for the univariate as well as for the multivariate case. The various results typically differ in the exact assumption for the dependence structure of the underlying time series data $X_1, \ldots, X_n$. Lu (2001) proved the asymptotic normality of a kernel density estimator under a so-called *near epoch* dependence, which is weaker than the notion of strong mixing (Definition 1) introduced above. Informally speaking, a sequence of observations is called near epoch dependent if a future observation can be approximated by an autoregressive function of increasing lag order.

In contrast to the independent and identically distributed and the above-described weakly dependent cases, one obtains for long range dependent observations that the asymptotic distribution of a kernel density smoother does not localize. It is only tight in the usual function spaces with supremum distances (Csörgö and Mielniczuk (1995)). Long-range dependence is said to be present if the autocovariance function of the time series is not summable or, equivalently, if the spectral density has a pole at zero frequency. For a recent application in the case of long range dependent volatility models, see also Casas and Gao (2008).

A further and promising dependence concept is the notion of *weak dependence*, which was introduced by Doukhan and Louhichi (1999). It has been successfully applied to different linear and especially nonlinear time series situations. The notion of weak dependence assumes bounds on the covariance of functions of observation tuples that are separated in time. Such an assumption typically can been verified in various nonlinear situations.

For the practical implementation of kernel density estimators, one major problem is the (data-adaptive) selection of a proper bandwidth $h > 0$. Results

for selection rules based on leave-one-out or cross-validation approaches can, for example, be found in Hall et al. (1995) or Hart and Vieu (1990).

For examples of applications of kernel density estimation in finance we refer to Stanton (1997), Donald (1997) and Aït-Sahalia (1996b). The latter-most paper, for example, applies nonparametric kernel density estimation in order to price interest rate securities. Pritsker (1998) investigated the performance of nonparametric density estimators (and also nonparametric regression) when applied to persistent time series.

2.2 Kernel smoothing regression

Having dealt with nonparametric density estimation for dependent data, we now proceed to the further topic of kernel and local polynomial smoothing in nonparametric regression. The main goal of nonparametric regression is the estimation of a conditional expectation $m(x) = E[Y|X = x]$ of a random variable Y given X on the basis of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

In many applications X and Y are observations of one and the same time series, e.g., $m(x) = E[X_t|X_{t-1} = x]$ represents the simplest univariate case or in a multivariate setup we may have $m(x) = E[X_t|(X_{t-1}, \ldots, X_{t-k}) = (x_1 \ldots, x_k)]$. A simple model in this framework is a nonparametric AR–ARCH model of the following form

$$X_t = m(X_{t-1}) + \sigma(X_{t-1}) \cdot \varepsilon_t \ , \ t \in \mathbb{Z} \tag{4}$$

with error variables ε_t that have conditional mean zero and unit variance, given the past $X_{t-1}, X_{t-2}, \ldots$. For nonparametric kernel smoothing in such a discrete time model see Franke et al. (2002a).

In the following, we describe the local polynomial estimator (LPE) for estimation of a conditional expectation. LPEs were introduced in Stone (1977). Tsybakov (1986), Fan (1992, 1993) and Fan and Gijbels (1995) discuss the behavior of LPEs for nonparametric regression in full detail. Masry (1996a) and Hjellvik and Tjøstheim (1995) applied LPEs to dependent data, including nonparametric autoregressive models.

Local polynomial smoothers are defined as follows. A pth-order LPE $\widehat{m}_h^p(x)$ of $m(x) = E[Y|X = x]$ is given as $\widehat{a}_0 = \widehat{a}_0(x, X_0, \ldots, X_T)$, where $\widehat{a} = (\widehat{a}_0, \ldots, \widehat{a}_{p-1})'$ minimizes

$$M_x = \sum_{t=1}^{T} K\left(\frac{x - X_{t-1}}{h}\right)\left(X_t - \sum_{q=0}^{p-1} a_q \left(\frac{x - X_{t-1}}{h}\right)^q\right)^2 . \tag{5}$$

Here, as above, $h > 0$ denotes the bandwidth h of the LPE. For $p = 0$ the local polynomial smoother is also called a local constant smoother or Nadaraya–Watson estimator.

Typically, one assumes that the kernel K is a nonnegative density function of bounded total variation with $\mathrm{supp}(K) \subseteq [-1,1]$. For Nadaraya–Watson smoothing $p = 0$, one also uses kernels that have vanishing higher-order moments. These kernel functions must also have negative values. For such kernels, Nadaraya–Watson smoothers achieve faster rates of convergence if the regression function m fulfills higher-order smoothness conditions. For positive kernels, the same rates can be achieved for smooth regression functions m by choosing higher values of p. For K, no smoothness assumptions beyond Lipschitz continuity are needed, at least they do not lead to a better asymptotic performance of the estimator. For finite samples, more regular kernels can result in smoother estimated regression functions.

A solution of the least-squares problem in (5) leads to the following representation of the LPE $\widetilde{m}_h^p$:

$$\begin{aligned} \widetilde{m}_h^p(x) &= \sum_{t=1}^{n} w_h(x, X_{t-1}, \{X_0, \ldots, X_{n-1}\}) X_t \\ &= \left[(D_x' K_x D_x)^{-1} D_x' K_x \underline{X} \right]_1 , \end{aligned} \tag{6}$$

where $\underline{X} = (X_1, \ldots, X_n)'$,

$$D_x \;=\; \begin{pmatrix} 1 & \frac{x-X_0}{h} & \cdots & (\frac{x-X_0}{h})^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x-X_{n-1}}{h} & \cdots & (\frac{x-X_{n-1}}{h})^{p-1} \end{pmatrix}$$

and

$$K_x \;=\; \mathrm{Diag}\left[K(\frac{x-X_0}{h}), \ldots, K(\frac{x-X_{n-1}}{h}) \right] \;.$$

Above, $[\cdot]_1$ denotes the first entry of a vector.

At first sight, the analysis of $\widetilde{m}_h$ seems to be quite involved, because the X_t's are dependent and enter into the right-hand side of (6) at several places. But typically, it can be shown that by some approximations one can replace the LPEs by quantities of much simpler structure. In a first step the weights of the LPE can be written as (cf. (6))

$$\begin{aligned} & w_h(x, X_{t-1}, \{X_0, \ldots, X_{n-1}\}) \\ &= \sum_{q=0}^{p-1} d_q(x, \{X_0, \ldots, X_{n-1}\}) K\left(\frac{x - X_{t-1}}{h}\right) \left(\frac{x - X_{t-1}}{h}\right)^q \;, \end{aligned}$$

where $d_q(x, \{X_0, \ldots, X_{n-1}\}) = ((D_x' K_x D_x)^{-1})_{1,q+1}$.

Here, M_{ij} denotes the (i,j)th entry of a matrix M. The functions d_q depend on $\{X_0, \ldots, X_{T-1}\}$ in a rather smooth manner. Under reasonable assumptions, $d_q(x, \{X_0, \ldots, X_{n-1}\})$ can be replaced asymptotically by

$d_q^{(\infty)}(x) = \left((ED'_x K_x D_x)^{-1}\right)_{1,q+1}$. In an asymptotic discussion this allows us to approximate $w_h(x, X_{t-1}, \{X_0, \ldots, X_{n-1}\})$ by the weights

$$\overline{w}(x, X_{t-1}) \;=\; \sum_{q=0}^{p-1} d_q^{(\infty)}(x) K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q \;,$$

which depend only on a single value X_{t-1} .

For the Nadaraya–Watson kernel regression estimator (i.e., $p = 0$) one gets

$$\widetilde{m}_h^0(x) = \frac{\sum_{t=1}^n K\left(\frac{x-X_t}{h}\right) Y_t}{\sum_{t=1}^n K\left(\frac{x-X_t}{h}\right)} \;. \tag{7}$$

Asymptotic results for LPEs are widely available in the literature. Typically and as mentioned above, it is assumed that the underlying observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ fulfill a strong mixing assumption. Under some further assumptions, Masry (1996a) gives a pointwise asymptotic (i.e., for fixed x) normality result for $\widetilde{m}_h^p(x)$. As an extension, Masry (1996b) deals with uniform asymptotic results (i.e., asymptotic results of the behavior of supremum distances of the LPE from the underlying function m). Moreover, the special case of univariate observations can be found in Masry and Fan (1997). For the special case where all covariables are lagged observations of the underlying time series itself, we refer to Masry and Tjøstheim (1994).

Nonparametric modeling for stochastic volatility (SV) models is also related to nonparametric regression and autoregression. A simple SV model is given by the following model equation:

$$X_t = \sigma_t \eta_t \;,\; \log(\sigma_t^2) = m\left(\log \sigma_{t-1}^2\right) + s\left(\log \sigma_{t-1}^2\right) \cdot \varepsilon_t \;,\; t = 1, \ldots, n. \tag{8}$$

Here, the volatility is driven by a nonparametric AR-ARCH model. The relation of this model to nonparametric autoregression is obvious. In (8) it is assumed that the errors η_t have conditional mean of 0 and conditional variance of 1, given the past, and that the conditional mean of ε_t is zero. Additionally, one can also assume that the bivariate errors (η_t, ε_t) are independent and identically distributed, allowing or not allowing a dependence between η_t and ε_t, e.g., a nonvanishing or vanishing correlation. Nonparametric inference for such models is discussed in Franke et al. (2003). As expected, owing to the fact that we only can observe the variables X_t, it turns out that the situation is related to nonparametric regression with errors in variables. The errors in observing the covariates complicate statistical inference and lead to slow convergence rates for nonparametric estimators. Related results for volatility density estimation are found in Van Es et al. (2003, 2005).

LPEs are not the only possible approach for nonparametric estimation of regression functions. Possible alternative methods are, on one hand, orthogonal series estimators and, on the other hand, smoothing or regression splines. We refer to Wahba (1990), Stone (1994), Newey (1997) and Härdle

et al. (1998). The reason why we mainly restrict our discussion within this chapter to kernel smoothing is that all these alternative methods typically do not allow for an explicit expression of the limiting distribution. This leads to some difficulties in the implementation of testing procedures and the construction of confidence intervals. For local polynomials we shall discuss tests and confidence intervals in subsequent sections.

One may consider the behavior of nonparametric procedures under dependency conditions other than strong mixing. Above, we already mentioned the concept of weak dependence introduced in Doukhan and Louhichi (1999). Asymptotic results of Nadaraya–Watson kernel type regression estimators under this dependence concept can be found in Ango Nze et al. (2002) and Ango Nze and Doukhan (2004). The concept of near epoch dependence (see above) was used by Lu and Linton (2007) to treat local linear fitting.

We conclude this section by briefly mentioning some applications of nonparametric density and regression estimates in the field of financial time series. Aït-Sahalia and Lo (1998) suggested a nonparametric fitting of option prices and moreover constructed a nonparametric estimator of the state-price density. Aït-Sahalia and Duarte (2003) also dealt with shape-constrained nonparametric estimation and moreover gave applications to finance. Fan and Yao (1998) applied local linear regression to the squared residuals for estimating the conditional variance or volatility function of stochastic regression with a focus on finance.

2.3 Diffusions

In the field of continuous-time modeling of financial data a time-homogeneous diffusion process $(X_t : t \geq 0)$ is frequently used. Such a process is given by the following stochastic differential equation:

$$dX_t = m(X_t)\,dt \;+\; \sigma(X_t)\,dW_t \;. \tag{9}$$

In (9) $(W_t : t \geq 0)$ denotes a standard Brownian motion and the functions $m(\cdot)$ and $\sigma^2(\cdot)$ are the drift (or mean) and the diffusion or variance process, respectively. We refer to Jiang and Knight (1997) for regularity conditions ensuring regular and stationary solutions of (9).

In a nonparametric approach one wants to estimate both functions without assuming any parametric functional form. Of course, especially an adequate specification of the diffusion function is of great importance because it has a great influence on pricing of derivatives.

A discrete version of (9) is given by

$$X_{t+\delta} - X_t = m(X_t) \cdot \delta + \sigma(X_t) \cdot \sqrt{\delta} \cdot \varepsilon_t \;, \tag{10}$$

where (ε_t) denotes a sequence of independent and identically distributed and standard normally distributed random variables. Equation (10) is known as the so-called *Euler-approximation* scheme with approximation rate $\delta^{-1/2}$. An alternative discrete-time model is obtained via the *Itô–Taylor* expansion with convergence rate δ^{-1}, which reads as follows

$$X_{t+\delta} - X_t = m(X_t) \cdot \delta + \sigma(X_t) \cdot \sqrt{\delta} \cdot \varepsilon_t + \frac{1}{2}\sigma^2(X_t) \cdot \delta \cdot (\varepsilon_t^2 - 1) \ . \quad (11)$$

The two functions m and σ^2 can be derived from the first two moments of the conditional distribution of increments

$$m(x) \ = \ \lim_{\delta \to 0} E\left(\frac{X_{t+\delta} - X_t}{\delta} \, | X_t = x \right) \quad (12)$$

and

$$\sigma^2(x) \ = \ \lim_{\delta \to 0} E\left(\frac{(X_{t+\delta} - X_t)^2}{\delta} \, | X_t = x \right) \ . \quad (13)$$

It is obvious how nonparametric regression estimators can be used in this context if we observe the diffusion X_t on a fine grid $t = k\delta, k = 1, \ldots, n$, of interval $[0, T]$. The Euler approximation suggests applying the classical nonparametric regression estimators with $X_k \equiv X_{k\delta}$ as regressor variables and $Y_k \equiv (X_{(k+1)\delta} - X_{k\delta})/\delta$ as responses. Proceeding as in (7), we obtain an estimator $\widehat{m}(x)$ of $m(x)$. Using $Z_k \equiv (X_{(k+1)\delta} - X_{k\delta})^2/\delta$ as responses, we get a nonparametric estimator $\widehat{\sigma}^2(x)$ of the diffusion function (Stanton (1997), Fan (2005)).

An asymptotic investigation can be done by letting $\delta \to 0$. Since the standard deviation of Y_k approximately is of order $\delta^{-1/2}$ we get nonconsistency for the estimator of the drift (in contrast to the estimation of the diffusion function). An alternative estimator of the drift function is given by the following equation (Jiang and Knight (1997)):

$$\widehat{m}(x) = \frac{1}{2\widehat{\pi}(x)} \frac{\partial \left(\widehat{\sigma}^2(x)\widehat{\pi}(x) \right)}{\partial x} \ . \quad (14)$$

Here, π denotes the stationary density (which is assumed to exist) and $\widehat{\pi}(x)$ is an estimator of $\pi(x)$, e.g., the kernel density estimator. Under some regularity conditions, Jiang and Knight (1997) showed pointwise consistency of the estimator (13) and derived its asymptotic distribution. Furthermore and under stronger conditions, consistency of the modified drift function estimator (14) is derived. Jiang and Knight (1997) also applied the proposed nonparametric estimators to short-term interest rate models.

Jacod (2000) investigated a nonparametric estimator of kernel type for the diffusion function in situations in which we observe our data at time points $k/n, k = 1, \ldots, n$. Hoffmann (2001) and Bandi and Nguyen (2003) provided a general asymptotic theory for nonparametric estimates of the

diffusion function in a stationary situation based on discrete observations. Hoffmann (2001) also gave optimal rates in the minimax sense. Mostly, the asymptotic analysis is done for an increasing sample size that comes from more frequent observations in a fixed interval, e.g., $k/n, k = 1, \ldots, n$. An alternative approach would allow an increasing observation period. This is covered in Bandi and Phillips (2003), who investigated discretely observed homogeneous stochastic differential equations with minimal requirements for the data generating process. The asymptotic analysis in this paper was done for increasing observation frequency as well as for increasing observation time span. The method developed applies also to nonstationary data.

Stanton (1997) suggested the use of higher-order approximations of conditional expectations like the mean and variance function in nonparametric diffusion. Stanton developed on the basis of these approximations nonparametric kernel estimates for continuous-time diffusion processes that are observed at discrete grids. He illustrated his results for various financial data sets. Fan and Zhang (2003) discussed the proposal of Stanton (1997) and investigated asymptotic bias and variance properties of the estimators. The main assumptions in order to obtain consistency and asymptotic expressions for bias and variance are the typical assumptions regarding the bandwidth h of the kernel estimator (like $h \to 0$ and $nh \to \infty$ or $nh^3 \to \infty$ as $n \to \infty$). Furthermore it is assumed that the distance δ between two subsequent observations of the diffusion process fulfills $\delta \to 0$ and that δ multiplied by some power of the sample size n converges to infinity. The last assumption ensures that the range over which one can (discretely) observe the diffusion increases.

A completely different asymptotic approach is needed for the discussion of *low-frequency observations*. Here, for discretely observed data $X_{k\cdot\delta}, k = 0, 1, \ldots, n$, the sampling frequency δ^{-1} is not large. This suggests an asymptotic approach where δ is fixed as the number of observations n tends to infinity. The estimation problem for $\sigma(\cdot)$ turns out to be more involved. The main idea in Gobet et al. (2004) is to construct in a first step an estimator of quantities of the form $E\left[f(X_\delta)\,|X_0 = x\right]$ and to identify in a second step from such conditional expectations the unknown variance function $\sigma(\cdot)$. It shows that this more or less leads to a deconvolution problem. For a thorough investigation of the situation of fixed sampling frequency we refer to Gobet et al. (2004). For an extension to nonparametric estimation for Lévy processes from low-frequency observations see Neumann and Reiss (2007).

3 Testing

A major motivation behind the application of nonparametric methods is checking and validating parametric and semiparametric models. Such a check can be based on a comparison of the statistical results from a nonparametric analysis with the parametric or semiparametric findings. Formally, this leads

to specification or goodness-of-fit tests. In this section we will discuss such approaches.

A correct specification as well as an adequate model evaluation is especially important for the variance, or volatility function, $\sigma(\cdot)$ because this function directly affects derivative pricing. This is true for discrete-time and continuous-time models. Testing whether or whether not a hypothetical continuous-time model is correctly specified has attracted much interest in the literature. A major part of the literature concentrates on the case of discretely observed samples $X_{t\delta}, t = 1, \ldots, n$, arising from discrete approximations like (10).

An early reference promoting the idea of comparing a completely nonparametric estimator and its model-based, typically parametric, counterpart is Härdle and Mammen (1993). In this paper, such an approach was discussed and investigated in detail for independent and identically distributed (X_t, Y_t). The authors used the regression problem as an example and mainly considered the comparison along the usual L_2 distance of both regression estimates. In the case of a parametric hypothesis for a regression function m, the test uses a parametric estimator $m_{\widehat{\theta}}$ and a nonparametric estimator, e.g., the Nadaraya–Watson estimator $\widetilde{m}_h^0$, defined in (7). In the paper, it is argued that it is not reasonable to use the L_2-distance test statistic $T_n = \int [\widetilde{m}_h^0(x) - m_{\widehat{\theta}}(x)]^2 w(x)\, dx$, where w is some weight function. We briefly explain this for the special case of a simple hypothesis $m = m_0$ and $m_{\widehat{\theta}} = m_0$. For this case one can show that (for certain choices of bandwidths) a test based on the statistic T_n is asymptotically equivalent to a linear test based on the statistic $2 \int [E_0^*[\widetilde{m}_h^0(x)] - m_0(x)]][\widetilde{m}_h^0(x) - E_0^*[\widetilde{m}_h^0(x)]] w(x)\, dx$. Here, $E_0^*[\cdot]$ denotes the conditional expectation on the hypothesis $m = m_0$, given the covariates X_t. Thus, asymptotically, the test T_n is not an omnibus test but a linear test that checks only for deviations in one direction. This misbehavior of the test is caused by the fact that the bias of the nonparametric estimator does not vanish, i.e., that $E_0^*[\widetilde{m}_h^0(x)] - m_0(x)$ is not asymptotically negligible. A simple idea solves this bias problem. If one adds to m (or in general to $m_{\widehat{\theta}}$) a term that is asymptotically equivalent to the bias of $\widetilde{m}_h^0$ on the hypothesis, then in the difference of $\widetilde{m}_h^0$ and the biased version of $m_{\widehat{\theta}}$ the bias cancels out. One way to add a bias is to use a smoothed version of m (or of $m_{\widehat{\theta}}$, respectively), i.e.,

$$\frac{\sum_{t=1}^n K\left(\frac{x - X_t}{h}\right) m_{\widehat{\theta}}(X_t)}{\sum_{t=1}^n K\left(\frac{x - X_t}{h}\right)} .$$

Implementing this new estimator of m into an L_2-distance test statistic results in the following test statistic:

$$\widetilde{T}_n = \int \left[\frac{\sum_{t=1}^n K\left(\frac{x - X_t}{h}\right) [Y_t - m_{\widehat{\theta}}(X_t)]}{\sum_{t=1}^n K\left(\frac{x - X_t}{h}\right)} \right]^2 w(x)\, dx.$$

This is the test statistic that was proposed in Härdle and Mammen (1993). There it was shown that $\widetilde{T}_n$ has an asymptotic normal distribution and that it behaves like an omnibus test. The asymptotic normality result is only good for theoretical considerations and it is not accurate enough for applications to data sets with moderate sample sizes. The reason is that asymptotic normality comes from asymptotic independence of values of the nonparametric estimator $\widetilde{m}_h^0$ at points that differ more than h. In a certain sense, one applies a central limit theorem for a sum of $\mathrm{O}(h^{-d})$ independent summands, where d is the dimension of the covariates X_t. Also for relatively large values of the sample size, the value of h^{-d} may be relatively small, leading to a poor normal approximation of the test statistic. In Härdle and Mammen (1993) the wild bootstrap idea from parametric linear models was proposed for nonparametric regression and its asymptotic consistency was shown. Kreiss et al. (2008) followed this proposal and considered testing problems for simple models in time series regression with applications to volatility testing. Besides parametric models they also considered lower-dimensional nonparametric models (e.g., additive nonparametric models) as hypothetical models to be tested.

Aït-Sahalia (1996a) applied the concept of parametric versus nonparametric fits in order to test a parametric structure of stationary densities, which are implied by (parametric) continuous-time models. Hong and Li (2002) applied this approach to testing parametric models for the transition densities. They did this in univariate diffusions and in various related and extended models of continuous-time financial time series. See also Chen et al. (2008) for a model specification test of a parametric diffusion process based on kernel estimation of the transitional density of the underlying process. All approaches can be viewed as goodness-of-fit tests (based on discrete observations) for specific continuous-time models.

Aït-Sahalia et al. (2001) gave a general description of these nonparametric testing methods and they applied them to testing different specifications of option prices, including the standard parametric Black–Scholes model, semiparametric specifications of the volatility surfaces and general nonparametric specifications. They proposed a test of a restricted specification of regression. Their test is based on comparing the residual sum of squares from kernel regression. They also discussed cases where both the restricted specification and the general model are nonparametric.

Following a comparable line of argument, Aït-Sahalia et al. (2005) developed a specification test for the existence of jumps in discretely sampled jump-diffusions, again based on a comparison of a nonparametric estimate of the transition density or distribution function with their corresponding parametric counterparts.

Further recent applications of the testing approach described to financial data are presented in Arapis and Gao (2006) and Thompson (2008).

Finally, we refer to Gao et al. (2007) for some general aspects of specification testing in cases where nonstationary covariables are present.

Model validation can also be based on confidence bands for the function of interest. A valid test, e.g., for a parametric hypothesis, is given by checking whether the parametric estimate is entirely contained within such a band. For the implementation of this idea the main problem is the computation of the (asymptotic) distribution of a kind of supremum distance of the nonparametric estimator from the true underlying function. For the simple model (4) corresponding results can be found in Neumann and Kreiss (1998), Franke et al. (2002b) and Zhao and Wu (2007). Especially, Franke et al. (2002b) gave consistency results for uniform bootstrap confidence bands of the autoregression function based on kernel estimates. Franke et al. (2004) extended those results to kernel estimates of the volatility function and applied them to checking for symmetry of volatility. The application of simultaneous confidence bands to nonparametric model validation in a variety of models used in financial econometrics can be found in Zhao (2008b).

4 Nonparametric Quantile Estimation

Conditional quantiles are the key to the most important risk measures for financial time series. The value at risk (VaR) is directly given by a quantile of the return distribution, but also the expected shortfall as a coherent measure of risk. Artzner et al. (1997) used conditional quantiles in its definition as expected excess loss above the VaR.

Let Y_t denote a stationary real-valued time series representing, e.g., the return of an asset at time t. Then, for given level α, the conditional quantile q_t^α of Y_t given information $\mathcal{F}_{t-1}$ is defined by

$$\mathrm{pr}(\, Y_t \le q_t^\alpha \mid \mathcal{F}_{t-1} \,) = \alpha.$$

In financial statistics, many methods for estimating the quantile q_t^α are volatility based, i.e., they start from a general stochastic volatility (SV) model

$$Y_t = \mu_t + \eta_t, \quad \eta_t = \sigma_t \epsilon_t$$

with independent and identically distributed (0,1) innovations ϵ_t, and with μ_t, σ_t^2 being the conditional mean and variance of Y_t given $\mathcal{F}_{t-1}$. Then, the conditional quantile is simply $q_t^\alpha = \mu_t + \sigma_t q_\epsilon^\alpha$, where q_ϵ^α denotes the α-quantile of the law of the innovations. For VaR calculations, a popular method is based on the GARCH(1,1) model, where $\mu_t = 0, \sigma_t^2 = \omega + aY_{t-1}^2 + b\sigma_{t-1}^2$.

A nonparametric approach may be based on a nonparametric autoregression with exogenous input and autoregressive conditional heteroscedasticity with exogenous components (ARX–ARCHX–like model), where $\mu_t = \mu(X_{t-1}), \sigma_t^2 = \sigma^2(X_{t-1})$ and where the d-dimensional predictor variable X_{t-1} may consist of finitely many past returns $Y_s, s < t$, as well as of past values of other financial time series, e.g., index returns, market trend and volatility

indicators, foreign exchange and interest rates, etc. To estimate the conditional quantile q_t^α in such a model, it suffices to estimate the local trend and volatility functions $\mu(x)$ and $\sigma(x)$, which is discussed in Section 2.2.

The volatility-based approach to estimating conditional quantiles has two drawbacks. On one hand, the distribution of the innovations ϵ_t has to be specified, e.g., as standard normal or as heavy-tailed like some t distribution. On the other hand, Engle and Manganelli (2004) have pointed out that volatility-based estimates of VaR tacitly assume that extreme negative returns show the same kind of random pattern as the majority of typical returns which mainly determine the volatility estimate. To avoid these problems, Engle and Manganelli (2004) considered a class of models, called CaViaR, i.e., conditional AR VaR, where the conditional quantile q_t^α is specified as a parametric function of finitely many of its own past values as well as of past returns. They proposed estimating the unknown model parameters directly by following the regression quantiles approach of Koenker and Bassett (1978); see also Koenker (2005). It is based on observing that the conditional α-quantile is given as

$$q_t^\alpha = \arg\min_{q\in\mathbb{R}} E\big\{|Y_t - q|_\alpha \;\big|\; \mathcal{F}_{t-1}\big\}, \tag{15}$$

where $|z|_\alpha = (\alpha-1)z1_{(-\infty,0]}(z) + \alpha z 1_{(0,\infty)}(z)$ denotes the so-called check function.

For estimating conditional quantiles directly in a nonparametric manner, it is convenient to replace the common nonparametric AR or ARX models by a quantile ARX model (QARX):

$$Y_t = q^\alpha(X_{t-1}) + \eta_t,$$

where the conditional quantile of η_t given $\mathcal{F}_{t-1}$ is zero. Then, $q_t^\alpha = q^\alpha(X_{t-1})$. If this is a model for asset returns Y_t and $\alpha \approx 0$, the quantile function $q^\alpha(x)$ directly describes the location of the extreme losses without having to consider some scale measure. If some quantile analogue of volatility is required nevertheless, the QARX model may be specified to a QARX–ARCHX process

$$Y_t = q^\alpha(X_{t-1}) + s^\alpha(X_{t-1})\epsilon_t,$$

where the quantile innovations ϵ_t are independent and identically distributed to α-quantile 0 and α-scale 1. Here, the α-scale of a real random variable Z having α-quantile q_Z^α is defined as the α-quantile of $|Z - q_Z^\alpha|_\alpha$.

Similarly to estimating the conditional mean by a local least-squares polynomial estimate, we get nonparametric estimates of the quantile function $q_\alpha(x)$ by minimizing the local sample version of (15). For the special case of a local constant approximation, we get a kernel estimate $\widehat{q}_\alpha(x,h)$ from a sample $(Y_t, X_{t-1}), t = 1,\ldots,N$, as

$$\widehat{q}_\alpha(x,h) = \arg\min_{q\in\mathbb{R}} \frac{1}{Nh}\sum_{t=1}^{N} \big|Y_t - q\big|_\alpha K\left(\frac{x - X_{t-1}}{h}\right),$$

where K is a kernel function as in Section 2. Equivalently, we get $\widehat{q}_\alpha(x,h)$ by inverting the corrresponding Nadaraya–Watson kernel estimate,

$$\widehat{F}_h(y|x) = \frac{\sum_{t=1}^{N} 1_{(-\infty,y]}(Y_t)\; K\big(\frac{x-X_{t-1}}{h}\big)}{\sum_{t=1}^{N} K\big(\frac{x-X_{t-1}}{h}\big)},$$

of the conditional distribution function $F(y|x) = E\big\{1_{(-\infty,y]}(Y_t)\big|X_{t-1} = x\big\}$.

The quantile kernel estimate $\widehat{q}_\alpha(x,h)$ has theoretical properties similar to those of the corresponding Nadaraya–Watson kernel estimate of the conditional mean, e.g., consistency for $N \to \infty, h \to 0, Nh \to \infty$, bias and variance expansions, asymptotic normality and uniform consistency (Abberger (1996), Cai (2002), Franke and Mwita (2003)). It is related to the local median of Truong and Stone (1992) corresponding to $\alpha = 0.5$ and the special case of a rectangular kernel; see also Boente and Fraiman (1995). It provides a robust alternative to the Nadaraya–Watson estimate as a conditional measure of location.

Other nonparametric approaches to estimating conditional quantiles are based on neural networks Chen and White (1999) or, more generally, sieve estimates (Franke et al. (2007)), Section 6 of this chapter), or on support vector machines (Christmann (2004)).

5 Advanced Nonparametric Modeling

The nonparametric regression and autoregression models of the last sections are more or less of the form *response = nonparametric function + noise*. In particular, in the case of multivariate covariates, the nonparametric methods behave poorly for moderate sample sizes and achieve only slow rates of convergence. This so-called curse of dimensionality can be circumvented by more structured models. A classic example is additive nonparametric models where the components of the multivariate covariate vector enter only additively into the regression. Such additive nonparametric models are an important tool in nonparametric regression. They allow a flexible modeling, are easy to interpret and there exist stable and reliable estimators. Thus, they avoid a lot of problems that are present if one uses a full dimensional nonparametric regression model. A classic estimation approach in additive modeling is the *backfitting* estimator (Hastie and Tibshirani (1991)). Mammen et al. (1999) have developed a closed asymptotic theory for a version of backfitting called smooth backfitting. Related models appear for panel and cross-section data where nonparametric approaches can be used to describe the dynamics of the individual time series or of underlying factors. Nonparametric autoregression

models for panels were used in Mammen et al. (2008) and in Linton et al. (2008). These estimators are also defined as solutions of empirical integral equations. For semiparametric generalized ARCH (GARCH) type approaches see also the chapter by Linton (2008) in this handbook.

A further approach uses nonparametric methods to reduce the dimension of the underlying data to finite-dimensional summaries and studies the dynamics of the data by applying classical methods from vector autoregression to the finite-dimensional vectors. This approach was applied in Borak et al. (2008), Fengler et al. (2007) and Brüggemann et al. (2008) for modeling the dynamics of implied volatility surfaces. Theoretical work includes the proof of the following oracle property for different models. A vector autoregression analysis based on the fitted finite-dimensional process leads asymptotically in first order to the same results as an analysis based on the unobserved true process. This property allows a simple and closed asymptotic distribution theory for the statistical analysis.

Nonparametric principal component analysis for implied volatility surfaces was used in Cont and da Fonseca (2002) and Benko et al. (2008). Connor et al. (2007) considered an additive model with time-varying scalar coefficients and applied this approach to the Fama–French model and to testing the capital asset pricing model.

Linton and Sancetta (2007) used kernel smoothing to estimate conditional distributions and expectations given an infinite past. Their estimator was based on smoothing with respect to an increasing number of lagged values.

Financial modeling often requires the inclusion of nonstationary components. Nonparametric smoothing for nonstationary diffusions was discussed in Bandi and Phillips (2003); see also the discussion in Section 2.3. Nonparametric regression with nonstationary covariates and nonparametric density estimation for nonstationary processes were also considered in Phillips and Park (1998), Moloche (2001) and Karlsen and Tjøstheim (2001). The approach in the last two papers was generalized to additive models in Schienle (2007). It turns out that for the additive model consistent nonparametric estimation is possible, even in situations in which the full dimensional nonparametric regression model cannot be estimated consistently. Moreover, nonparametric regression models with nonstationary errors were studied in Linton and Mammen (2008). They used the method of differencing the nonparametric model in order to transform the data to an additive model with stationary errors.

Recently there has been growing interest in statistics of models, which allow for a huge number of parameters, even much larger than the sample size. The essential assumption is that a much smaller number of parameters is significantly different from zero. Data of this type naturally arise in finance. For a recent application with high-dimensional covariance matrices see Fan et al. (2007b). It also can be expected that the new developments in statistical learning with high-dimensional data will find further important applications in finance.

In Fan et al. (2007a), it is proposed to combine two independent nonparametric estimators of the volatility matrix—one based on the time domain and the other one based on the state domain—to form a new aggregated estimator in order to circumvent the curse of dimensionality.

Often, it is known that a function fulfills certain shape constraints such as monotonicity, convexity, etc. Such constraints naturally arise in different settings of finance. Incorporating the constraints in the estimation leads to a more accurate estimate. A general smoothing framework for nonparametric estimation under shape constraints is presented in Mammen (1991) and Mammen et al. (2001). Discussion of shape-constrained nonparametric estimation with applications to finance can be found in Aït-Sahalia and Duarte (2003).

6 Sieve Methods

Up to now, we have mainly considered nonparametric function estimates based on local smoothing. Owing to the local sparseness of data in spaces of higher dimensions, these methods work well only for low-dimensional predictor variables. To cope with this curse of dimensionality, semiparametric models or models with constraints like additivity on the functions to be estimated may be considered; see also the last section. Another possibility is to consider Grenander's method of sieves (Grenander (1981)). We illustrate the idea for a nonparametric ARX model

$$Y_t = m(X_{t-1}) + \epsilon_t,$$

where Y_t is real-valued and $X_{t-1} \in \mathbb{R}^d$ may consist of past values $Y_{t-1}, \ldots,$ Y_{t-p} as well as of past values of other time series. $(Y_t, X_{t-1}), \infty < t < \infty$, is assumed to be a stationary process. To estimate the autoregressive function $m(x)$ from a sample $(Y_t, X_{t-1}), t = 1, \ldots, n$, we approximate it by a function $m_n \in \mathcal{G}_n$, where $\mathcal{G}_n = \{g(x;\theta);\, \theta \in \Theta_n\}$ is a parametric class of functions, and the dimension D_n of the set Θ_n of admissible parameters depends on sample size n. We consider the nonlinear least-squares estimate of θ,

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta_n} \sum_{t=1}^{n} \big(Y_t - g(X_{t-1};\, \theta)\big)^2, \tag{16}$$

and we get $m_n(x) = g(x; \widehat{\theta}_n)$ as a sieve estimate of $m(x)$.

The method becomes nonparametric by choosing the function classes $\mathcal{G}_n$ to be increasing with sample size n and by assuming that they have a universal approximation property, i.e., their union $\mathcal{G}_\infty = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots$ is dense in an appropriate function space containing $m(x)$, e.g., the space $L^2(\mu)$ of functions that are square-integrable with respect to the stationary distribution of X_t.

Under appropriate assumptions regarding the time series and the estimation procedure, it can be shown that the sieve estimate is nonparametrically consistent in the sense that $E((m_n(X_s)-m(X_s))^2 = \int (m_n(x)-m(x))^2\mu(dx) \to 0$ for $n \to \infty$ provided that $D_n \to \infty$ too with the right rate (Franke and Diagne (2006)).

Frequently, the size of the parameters is bounded, i.e., $|\theta| \le \Delta_n, \theta \in \Theta_n$, for a kind of norm $|\cdot|$, and, then, the consistency depends on a combination of rate conditions on D_n and $\Delta_n \to \infty$. In practice, one could use a few large parameter values or, alternatively, many small parameters to approximate $m(x)$ without overfitting. Parameter dimension and size D_n and Δ_n determine the smoothness of the nonparametric function estimate similarly to the bandwidth in local smoothing. They can be data adaptively chosen by the same kind of methods (using cross-validation, bootstrap or other approximations of the mean-square prediction error), but sieve estimates are typically applied in situations where X_t is high-dimensional and the sample size n is large. In this case, an affordable and computationally feasible method is simple validation, i.e., part of the sample, say, $(Y_t, X_{t-1}), t = 1, \ldots, N$, is used as a training sample for calculating estimates for various parameter dimensions, and the rest of the data $(Y_t, X_{t-1}), t = N+1, \ldots, n$, are used as a validation sample for comparing the predictive performance of the resulting function estimates.

For analyzing financial time series, the still most popular type of sieve estimate uses output functions of feedforward neural networks. The theoretical properties of those estimates have been thoroughly investigated by White and his coworkers. Here, the corresponding function classes $\mathcal{G}_n$ consist of functions

$$g(x;\,\theta) = v_0 + \sum_{h=1}^{H_n} v_h \psi\big(w_{0i} + \sum_{i=1}^{d} w_{hi} x_i\big), \quad \sum_{h=0}^{H_n} |v_h| \le \Delta_n. \tag{17}$$

The $D_n = 1 + H_n + H_n(1+d)$-dimensional parameter vector θ consists of all the network weights $v_0, ..., v_{H_n}, w_{hi}, h = 1, ..., H_n, i = 0, ..., d$. ψ is a given function, for application to financial time series typically of sigmoid shape, e.g., $\psi(u) = \tanh(u)$.

For more than one decade, neural networks have been a well-established tool for solving classification and forecasting problems in financial applications (Bol et al. (1996), Refenes et al. (1996)). In particular, they have been used for forecasting financial time series to generate trading signals for the purpose of portfolio management as in Evans (1997) and Franke (1998). Alternatively, as for portfolio allocation, the main goal is to achieve on average a large return combined with a small risk, not to get precise forecasts of future prices. Neural networks may be used to get the allocation directly as the output of a network by replacing the empirical mean-square prediction error in (16) by an economically meaningful performance measure, e.g., the risk-adjusted return (Heitkamp (1996), Franke and Klein (1999)). In risk management, neural networks were used for nonparametric estimation of volatility in

Franke (2000) and Franke and Diagne (2006) and of corresponding volatility-based VaR (see Section 4). If, for example, a time series Y_t follows a nonparametric ARCHX model $Y_t = \sigma(X_{t-1})\epsilon_t$, then a neural-network-based volatility estimate is $\sigma_n(X_{t-1}) = g^{1/2}(X_{t-1}; \widehat{\theta}_n)$, where $g(x; \theta)$ is given by (17) and $\widehat{\theta}_n$ minimizes

$$\sum_{t=1}^{n} \big(Y_t^2 - g(X_{t-1}; \theta)\big)^2,$$

as $\sigma^2(x)$ is the conditional variance of Y_t given $X_{t-1} = x$.

Neural networks (Chen and White (1999)) and other sieve estimates (Franke et al. (2007)) can also be used for estimating conditional quantile functions directly, providing an approach to VaR calculation based on more information than just the past prices of the asset under consideration. Again, a regression quantile approach may be used (see Section 4). The quantile function $q^\alpha(x)$ is estimated by $q_n^\alpha(x) = g(x; \widehat{\theta}_n)$ with now

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta_n} \sum_{t=1}^{n} \big|Y_t - g(X_{t-1}; \theta)\big|_\alpha.$$

Another example of sieve estimates corresponds to the qualitative threshold ARCH (QTARCH) models of Gouriéroux and Montfort (1992); see also Section 5.4 of Gouriéroux (1997). Allowing for a general predictor variable, they are of the form

$$Y_t = \sum_{i=1}^{p} \alpha_i 1_{A_i}(X_{t-1}) + \left(\sum_{i=1}^{p} \beta_i 1_{B_i}(X_{t-1}) \right) \epsilon_t,$$

where ϵ_t are i.i.d. (0,1) innovations, and $A_1, \ldots, A_p$, $B_1, \ldots, B_p$ are two partitions of the predictor space $\mathbb{R}^d$. If p is allowed to increase with sample size n this model can be interpreted as a sieve approximation to the general nonparametric ARX–ARCHX model $Y_t = m(X_{t-1}) + \sigma(X_{t-1})\epsilon_t$ with classes $\mathcal{G}_n$ consisting of piecewise constant functions. Franke et al. (2007) considered quantile sieve estimates based on that function class including a classification and regression tree (CART) like algorithm (Breiman et al. (1984)) for choosing the partition adaptively from the data. In a related approach, Audrino and Bühlmann (2001) considered VaR calculation based on piecewise parametric GARCH models. Many other function classes which may be used for constructing sieve estimates are discussed in the survey of Györfy et al. (2002).

Sieve estimates have also been used for modeling the nonparametric components in semiparametric models. The efficiency for the resulting parametric estimators was considered in Chen and Shen (1998) and Ai and Chen (2003).

References

Abberger, K. (1996): *Nichtparametrische Schaetzung bedingter Quantile in Zeitreihen. Mit Anwendung auf Finanzmarktdaten.* Hartung–Gore, Konstanz.

Ai, C. and Chen, X. (2003): Efficient estimators of models with conditional moment restrictions containing unknown functions. *Econometrica* **71**, 1795–1843.

Aït-Sahalia, Y. (1996a): Testing continuous-time models of the spot interest rate. *Review of Financial Studies* **9**, 385–426.

Aït-Sahalia, Y. (1996b): Nonparametric pricing of interest rate derivative securities. *Econometrica* **64**, 527–560.

Aït-Sahalia, Y., Duarte, J. (2003): Nonparametric option pricing und shape restrictions. *Journal of Econometrics* **116**, 9–47.

Aït-Sahalia, Y., Lo, A.W. (1998): Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance* **53**, 499–547.

Aït-Sahalia, Y., Bickel, P.J., Stoker, T.M. (2001): Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics* **105**, 363–412.

Aït-Sahalia, Y., Fan, J. and Peng, H. (2005): Nonparametric transition-based tests for jump-diffusions. *Preprint.* Available at SSRN: http://ssrn.com/abstract=955820

Ango Nze, P.A., Doukhan, P. (2004): Weak dependence: Models and applications in econometrics. *Econometric Theory* **20**, 995–1045.

Ango Nze, P.A., Bühlmann, P., Doukhan, P. (2002): Weak dependence beyond mixing and asymptotics for nonparametric regression. *Annals of Statistics* **30**, 397–430.

Arapis, M. and Gao, J. (2006): Empirical comparisons in short-term interest rate models using nonparametric methods. *Journal of Financial Econometrics* **4**, 310–345.

P. Artzner, P., Delbaen, F., Eber, F.-J. and Heath, D. (1997): Thinking Coherently. *Risk Magazine* **10**, 68–71.

Audrino, F. and Bühlmann, P. (2001): Tree-structured GARCH models. *Journal of the Royal Statistical Society Series B* **63**, 727–744.

Bandi, F. and Nguyen, T. H. (2000): Fully nonparametric estimators for diffusions: Small sample analysis. *Working paper Graduate School of Business, The University of Chicago.*

Bandi, F. M. and Nguyen, T. H. (2003): On the functional estimation of jump-diffusion models. *Journal of Econometrics* **116**, 293–328.

Bandi, F. and Phillips, P.C.B. (2003): Fully nonparametric estimation of scalar diffusion models. *Econometrica* **71**, 241–283.

Benko, M., Härdle, W. and Kneip, A. (2008): Common functional principal components. *Annals of Statistics* to appear.

Boente, G. and Fraiman, R. (1995). Asymptotic distribution of smoothers based on local means and local medians under dependence. *Journal of Multivariate Analysis* **54**, 77–90.

Bol, G., Nakhaeizadeh, G. and Vollmer, K.-H. (Eds.) (1996): *Finanzmarktanalyse und -prognose mit innovativen quantitativen Verfahren.* Physica, Heidelberg.

Borak, S., Härdle, W., Mammen, E. and Park, B. (2008): Time series modelling with semiparametric factor dynamics. Preprint.

Bosq, D. (1996): Bosq, D. (1996): *Nonparametric statistics for stochastic processes: estimation and prediction. Lecture Notes in Statistics* **110**. Springer, New York.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984): *Classification and Regression Trees.* Wadsworth, Belmont.

Brüggemann, R., Härdle, W., Mungo, J. and Trenkler, C. (2008): VAR modeling for dynamic semiparametric factors of volatility strings. *Journal of Financial Econometrics* **6**, 361–381.

Cai, Z. (2002): Regression quantile for time series. *Econometric Theory* **18**, 169–192.

Cai, Z. and Hong, Y. (2003): Nonparametric methods in continuous-time fiance: A selective review. *In: Akritas, M. G. and Politis, D. N. (Eds.): Recent Advances and Trends in Nonparametric Statistics*, 283–302. Elsevier, Amsterdam.

Casas, I. and Gao, J. (2008): Econometric estimation in long-range dependent volatility models: Theory and practice. *Journal of Econometrics* to appear.

Chen, S. X., Gao, J. and Tang, C. (2008): A test for model specification of diffusion processes. *Annals of Statistics* **36**, 167–198.

Chen, X. and Shen, X. (1998): Sieve extremum estimates for weakly dependent data. *Econometrica* **66**, 289–314.

Chen, X. and White, H. (1999): Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory* **45**, 682–691.

Christmann, A. (2004): An approach to model complex high-dimensional insurance data. *Allgemeines Statistisches Archiv* **88**, 375–396.

Collomb, G. (1984): Propriétés de convergence presque complète du predicteur à noyau. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **66**, 441–460.

Connor, G., Hagmann, M. and Linton, O.B. (2007): Efficient estimation of a semiparametric characteristic-based factor model of security returns. *Preprint.*

Cont, R. and da Fontseca, J. (2002): The dynamics of implied volatility surfaces. *Quantitative Finance* **2**, 45–60.

Csörgö, S. and Mielniczuk, J. (1995): Density estimation under long-range dependence *Annals of Statistics* **23**, 990–999.

Donald, S.G. (1997): Inference concerning the number of factors in a multivariate nonparametric relationship. *Econometrica* **65**, 103–131.

Doukhan, P., Louhichi, S. (1999): A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and Their Applications* **84**, 313–342.

Dürkes, A. and Kreiss, J.-P. (2006): Weak dependence of nonparametric GARCH-models. *Technical report, TU Braunschweig.*

Engle, R.F. and Manganelli, S. (2004): CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics* **22**, 367–381.

Evans, O. (1997): Short-term currency forecasting using neural networks. *ICL Systems Journal* **11**, 1–17.

Fan, J. (1992): Design-adaptive nonparametric regression. *Journal of the American Statistical Association* **87**, 998–1004.

Fan, J. (1993): Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics* **21**, 196–216.

Fan, J. (2005): A selective overview of nonparametric methods in financial econometrics (with discussion). *Statistical Science* **20**, 317–357.

Fan, J. and Gijbels, I. (1995): *Local Polynomial Modelling and Its Applications. Theory and Methodologies.* Chapman and Hall, New York.

Fan, J. and Guo, J. (2003): Semiparametric estimation of value at risk. *Econometrics Journal* **6**, 261–290.

Fan, J. and Yao, Q. (1998): Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.

Fan, J. and Yao, Q. (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

Fan, J. and Zhang, C.(2003): A re-examination of Stanton's diffusion estimations with applications to financial model validation. *Journal of the American Statistical Association* **98**, 118–134.

Fan, J., Fan, Y. and Lv, J. (2007a): Aggregation of nonparametric estimators for volatility matrix. *Journal of Financial Econometrics* **5**, 321–357.

Fan, J., Fan, Y. and Lv, J. (2007b): High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* to appear.

Fengler, M., Härdle, W. and Mammen, E. (2007): Implied volatility string dynamics. *Journal of Financial Econometrics* **5**, 189–218.

Franke, J. (1998): Nonlinear and nonparametric methods for analyzing financial time series. *In: Kall, P. and Luethi, H.-J. (eds.): Operation Research Proceedings 98*. Springer, Heidelberg.

Franke, J. (2000): Portfolio management and market risk quantification using neural networks. *In: Chan, W.S., Li, W.K. and Tong, H. (Eds.): Statistics and Finance: An Interface*. Imperial College Press, London.

Franke, J. and Diagne, M. (2006): Estimating market risk with neural networks. *Statistics and Decisions* **24**, 233–253.

Franke, J. and Klein, M. (1999): Optimal portfolio management using neural networks—a case study. *Report in Wirtschaftsmathematik* **49**. University of Kaiserslautern.

Franke, J. and Mwita, P. (2003): Nonparametric estimates for conditional quantiles of time series. *Report in Wirtschaftsmathematik* **87**, University of Kaiserslautern.

Franke, J., Kreiss, J.-P., Mammen, E. (2002a): Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli* **8**, 1–37.

Franke, J., Kreiss, J.-P., Mammen, E., Neumann, M.H. (2002b): Properties of the nonparametric autoregressive bootstrap. *Journal of Time Series Analysis* **23**, 555–585.

Franke, J., Härdle, W. and Kreiss, J.-P. (2003): Nonparametric estimation in a stochastic volatility model. *In: Akritas, M. G. and Politis, D. N. (Eds.): Recent Advances and Trends in Nonparametric Statistics*, 302–313. Elsevier, Amsterdam.

Franke, J., Neumann, M.H. and Stockis, J.-P. (2004): Bootstrapping nonparametric estimates of the volatility function. *Journal of Econometrics* **118**, 189–218.

Franke, J., Stockis, J.-P. and Tadjuidje, J. (2007): Quantile sieve estimates for time series. *Report in Wirtschaftsmathematik* **105**, University of Kaiserslautern.

Gao, J. (2007): *Nonlinear Time Series: semiparametric and nonparametric methods. Monographs on Statistics and Applied Probability* **108**. Chapman and Hall, London.

Gao, J., King, M. L., Lu, Z. and Tjøstheim, D. (2007): Specification Testing in Nonlinear Time Series. *Preprint, School of Economics, The University of Adelaide*.

Gobet, E., Hoffmann, M. and Reiss, M. (2004): Nonparametric estimation of scalar diffusions based on low-frequency data. *Annals of Statistics* **32**, 2223–2253.

Gouriéroux, C. (1997): *ARCH Models and Financial Applications*. Springer, Heidelberg.

Gouriéroux, C. and Montfort, A. (1992): Qualitative threshold ARCH models. *Journal of Econometrics* **52**, 159–199.

Grenander, U. (1981): *Abstract Inference*. Wiley, New York.

Györfy, L., Kohler, M., Krzyzak, A. and Walk, H. (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer, Heidelberg.

Hall, P., Lahiri, S.N. and Truong, Y.K. (1995): On bandwidth choice for density estimation with dependent data. *Annals of Statistics* **23**, 2241–2263.

Härdle, W. and Linton, O. (1994): Applied nonparametric methods. *In: Engle, R. and McFadden, D. (Eds.): Handbook of Econometrics* **IV**. North-Holland, Amsterdam.

Härdle, W. and Mammen, E. (1993): Comparing nonparametric versus parametric regression fits. *Annals of Statistics* **21**, 1926–1947.

Härdle, W. and Tsybakov, A.B. (1997): Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics* **81**, 223–242.

Härdle, W., Lütkepohl, H. and Chen, R. (1997): A review of nonparametric time series analysis. *International Statistical Review* **65**, 49–72.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov (1998): *Wavelets approximation and statistical applications. Springer Lecture Notes in Statistics* **129**. Springer, New York.

Hart, J.D. (1996): Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* **6**, 115–142.

Hart, J.D. and Vieu, P. (1990): Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics* **18**, 873–890.

Hastie, T. and Tibshirani, R. (1991): *Generalized Additive Models*. Chapman and Hall, London.

Heitkamp, D. (1996): Methodische Aspekte bei der Entwicklung von Tradingmodellen auf der Basis Neuronaler Netze. *Wirtschaftsinformatik* **38**, 238–292.

Hjellvik, V. and Tjøstheim, D. (1995): Nonparametric tests of linearity for time series. *Biometrika* **82**, 351–368.

Hoffmann, M. (2001): On estimating the diffusion coefficient: Parametric versus nonparametric. *Annales de l'Institut Henri Poincaré* **37**, 339–372.

Hong, Y. and Li, H. (2002): Nonparametric specification testing for continuous time models with applications to term structure of interest rates. *Review of Financial Studies* **18**, 37–84.

Jacod J. (2000): Non-parametric kernel estimation of the diffusion coefficient of a diffusion. *Scandinavian Journal of Statistics* **27**, 83–96.

Jiang, G.J. and Knight, J.L. (1997): A nonparametric approach to the estimation of diffusion processes, with application to a short term interest model. *Econometric Theory* **13**, 615–645.

Karlsen, H. and Tjøstheim, D. (2001): Nonparametric estimation in null-recurrent time series. *Annals of Statistics* **29**, 372–416.

Koenker, R. (2005). *Quantile Regression* Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978): Regression quantiles. *Econometrica* **46**, 33–50.

Kreiss, J.-P. (2000): Nonparametric estimation and bootstrap for financial time series. *In: Chan, W.S., Li, W.K. and Tong, H. (Eds.): Statistics and Finance: An Interface.* Imperial College Press, London.

Kreiss, J.-P., Neumann, M.H. and Yao, Q. (2008): Bootstrap tests for simple structures in nonparametric time series regression. *Statistics and Its Interface* to appear.

Linton, O. B. (2008): Semiparametric and nonparametric ARCH modelling. *In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.): Handbook of Financial Time Series*,156–167. Springer, New York.

Linton, O. B. and Mammen, E. (2005): Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* **73**, 771–836.

Linton, O. B. and Mammen, E. (2008): Nonparametric transformation to white noise. *Journal of Econometrics* **142**, 241–264.

Linton, O. B. and Sancetta, A. (2007): Consistent estimation of a general nonparametric regression function in time series. *Preprint, The London School of Economics.*

Linton, O. B., Mammen, E., Nielsen, J. P. and Tanggaard, C. (2001): Estimating yield curves by kernel smoothing methods. *Journal of Econometrics* **105**, 185–223.

Linton, O., Nielsen, J. P. and Nielsen, S. F. (2008): Nonparametric regression with a latent time series. *Preprint.*

Lu, Z. (2001): Asymptotic normality of kernel density estimators under dependence. *Annals of the Institute of Statistical Mathematics* **53**, 447–468.

Lu, Z., Linton, O. (2007): Local linear fitting under near epoch dependence. *Econometric Theory* **23**, 37–70.

Mammen, E. (1991): Estimating a smooth monotone regression function. *Annals of Statistics* **19**, 724–740.

Mammen, E., Linton, O. B. and Nielsen, J. P. (1999): The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27**, 1443–1490.

Mammen, E., Marron, J.S., Turlach, B.A., Wand, M.P. (2001): A general projection framework for constrained smoothing. *Statistical Science* **16**, 232–248.

Mammen, E., Stove, B. and Tjøstheim, D. (2008): Nonparametric additive models for panels of time series. *Preprint.*

Masry, E. (1996a): Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and Their Applications* **65**, 81–101 [Correction (1997). **67**, 281].

Masry, E. (1996b): Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis* **17**, 571–599.

Masry, E. and Fan, J. (1997): Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*. **24**, 165–179.

Masry, E. and Tjøstheim, D. (1994): Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory* **11**, 258–289.

Meir, R. (2000): Nonparametric time series prediction through adaptive model selection. *Machine Learning* **39**, 5–34.

Modha, D.S. and Masry, E. (1996): Minimum complexity regression estimation with weakly dependent observations. *IEEE Transacvtions on Information Theory* **42**, 2133–2145.

Modha, D.S. and Masry, E. (1998): Memory-universal prediction of stationary random processes. *IEEE Transactions on Infomation Theory* **44**, 117–133.

Moloche, G. (2001): Kernel regression for non-stationary harris-recurrent processes. *MIT working paper*.

Neumann, M.H. and Kreiss, J.-P. (1998): Regression-type inference in nonparametric autorregression. *Annals of Statistics* **26**, 1570–1613.

Neumann, M.H. and Reiss, M. (2007): Nonparametric estimation for Lévy processes from low-frequency observations. *Preprint, arXiv:0709.2007v1 [math.ST]*.

Newey, W.K. (1997): Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**, 147–168.

Pagan, A, and Ullah, A. (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Phillips, P.C.B. and Park, J.Y. (1998): Nonstationary density estimation and kernel autoregression. *Cowles Foundation Discussion Paper* **1181**.

Pritsker, M. (1998): Nonparametric density estimation and tests of continuous time interest rate models. *Review of Financial Studies* **11**, 449–487.

Refenes, A.-P., Burgess, A.N. and Bentz, Y. (1996): Neural networks in financial engineering: A study in methodology. *IEEE Transactions on Neural Networks* **8**, 1222–1267.

Robinson, P.M. (1983): Nonparametric Estimators for Time Series. *Journal of Time Series Analysis* **4**, 185–207.

Schienle, M. (2007): Nonparametric nonstationary regression. *Preprint, Department of Economics, Universität Mannheim*.

Stanton, R. (1997): A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance* **52**, 1973–2002.

Stone, C.J. (1977): Consistent nonparametric regression. *Annals of Statistics* **5**, 595–620.

Stone, C.J. (1994): The use of polynomial splines and their tensor products in mutlivariate function estimation (with discussion). *Annals of Statistics* **22**, 118–184.

Tjøstheim, D. (1994): Non-linear time series: A selective review. *Scandinavian Journal of Statistics* **21**, 97–130.

Thompson, S. (2008): Identifying term structure volatility from the LIBOR–swap curve. *Review of Financial Studies* **21**, 819–854.

Tong, H. (1990): *Non-linear Time Series. A Dynamic System Approach*. Oxford University Press, Oxford.

Tsybakov, A.B. (1986): Robust reconstruction of functions by the local approximation method. *Problems of Information Transmission* **22**, 133–146.

Truong, Y.K. and Stone, C.J. (1992): Nonparametric function estimation involving time series. *Annals of Statistics* **20**, 77–97.

Van Es, B., Spreij, P. and van Zanten, H. (2003): Nonparametric volatility density estimation. *Bernoulli* **9**, 451–465.

Van Es, B., Spreij, P. and van Zanten, H. (2005): Nonparametric volatility density estimation for discrete time models. *Nonparametric Statistics* **17**, 237–251.

Whaba, G. (1990): *Spline models for observational data. CBMS-NSF Regional Conference Series in Apllied Mathematics* **59**. Society for Industrial and Applied Mathematics, Philadelphia.

White, H. and Wooldridge, W. (1990): Some results for sieve estimation with dependent observations. *In: Barnett, W., Powell, J. and Tauchen, G. (Eds.): Nonparametric and Semi-Parametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge..

Yang, L., Härdle, W. and Nielsen, J. (1999): Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis* **20**, 579–604.

Zhao, Z. (2008a): Parametric and nonparametric models and methods in financial econometrics. *Statistics Surveys, arXiv:0801.1599v1 [stat.ME]*.

Zhao, Z. (2008b): Nonparametric model validation for hidden Markov models with application to financial econometrics. *Preprint, Department of Statistics, Pennsylvania State University*.

Zhao, Z. and Wu, W. B. (2007): Confidence bands in nonparametric time series regression. *Annals of Statistics* **36**, 1854–1878.

Modelling Financial High Frequency Data Using Point Processes

Luc Bauwens and Nikolaus Hautsch

Abstract We survey the modelling of financial markets transaction data characterized by irregular spacing in time, in particular so-called financial durations. We begin by reviewing the important concepts of point process theory, such as intensity functions, compensators and hazard rates, and then the intensity, duration, and counting representations of point processes. Next, in two separate sections, we review dynamic duration models, especially autoregressive conditional duration models, and dynamic intensity models (Hawkes and autoregressive intensity processes). In each section, we discuss model specification, statistical inference and applications.

1 Introduction

Since the seminal papers by Hasbrouck (1991) and Engle and Russell (1998) the modelling of financial data at the transaction level is an ongoing topic within financial econometrics. This has created a new literature, often referred to as "the econometrics of (ultra-)high-frequency finance" or "high-frequency econometrics". The peculiar properties of financial transaction data, such as the irregular spacing in time, the discreteness of price changes, the bid-ask bounce as well as the presence of strong intraday seasonalities and persistent dynamics, has spurred a surge in new econometric approaches. One important strand of the literature deals with the irregular spacing of data in time. Taking into account the latter is indispensable if one seeks to exploit the full amount

Luc Bauwens
Université catholique de Louvain, CORE, Voie du Roman Pays 34, 1348, Louvain-la-Neuve, Belgium, e-mail: luc.bauwens@uclouvain.be

Nikolaus Hautsch
Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics, Spandauer Str. 1, D-10099 Berlin, Germany, e-mail: nikolaus.hautsch@wiwi.hu-berlin.de

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_41,

of information in financial transaction data. For example, the timing and frequency of order and trade arrivals carry information on the state of the market and play an important role in market microstructure analysis, for the modelling of intraday volatility as well as the measurement and prediction of liquidity, transaction costs and implied liquidity risks.

Accounting for the irregular occurrence of transaction data requires us to model the series as so-called *financial point processes*. We label the inter-event waiting times as financial durations and classify them according to the events of interest, with the most common being *trade durations* and *quote durations*, defined by the time between two consecutive trade or quote arrivals, respectively. *Price durations* correspond to the time between cumulative absolute price changes of given size and can be used as a volatility measure. Similarly, a *volume duration* is defined as the time until a cumulative order volume of given size is traded and captures an important dimension of market liquidity. For more details and illustrations, see Bauwens and Giot (2001) or Hautsch (2004).

One important property of transaction data is that market events are clustered over time implying that financial durations follow positively autocorrelated processes with strong persistence. In fact, the dynamic properties of financial durations are quite similar to those of (daily) volatilities. These features may be captured in alternative ways through different dynamic models based on either duration, intensity or counting representations of a point process.

This chapter reviews duration- and intensity-based models of financial point processes. In Section 2, we introduce the fundamental concepts of point process theory and discuss major statistical tools. In Section 3, we review the class of *dynamic duration models*. Specifying a (dynamic) duration model is arguably the most intuitive way to characterize a point process in discrete time and has been suggested by Engle and Russell (1998), which inspired a large literature. Nevertheless, Russell (1999) realized that a continuous-time setting on the basis of the intensity function constitutes a more flexible framework which is particularly powerful for the modelling of multivariate processes. Different types of *dynamic intensity models* are presented in Section 4.

2 Fundamental Concepts of Point Process Theory

In this section, we discuss important concepts in point process theory which are needed throughout this chapter. In Section 2.1, we introduce the notation and basic definitions. The fundamental concepts of intensity functions, compensators and hazard rates are defined in Section 2.2, whereas in Section 2.3 different classes and representations of point processes are discussed. Finally, in Section 2.4, we present the random time change theorem which yields a

powerful result for the construction of diagnostics for point process models. Most concepts discussed in this section are based upon Chapter 2 of Karr (1991).

2.1 Notation and definitions

Let $\{t_i\}_{i\in\{1,\ldots,n\}}$ denote a random sequence of increasing event times $0 < t_1 < \ldots < t_n$ associated with an orderly (simple) point process. Then, $N(t) := \sum_{i\geq 1} \mathbb{1}_{\{t_i\leq t\}}$ defines the right-continuous (càdlàg) counting function. Throughout this chapter, we consider only point processes which are integrable, i.e. $\mathrm{E}[N(t)] < \infty \; \forall\, t \geq 0$. Furthermore, $\{W_i\}_{i\in\{1,\ldots,n\}}$ denotes a sequence of $\{1,\ldots,K\}$-valued random variables representing K different types of events. Then, we call the process $\{t_i, W_i\}_{i\in\{1,\ldots,n\}}$ an K-variate *marked* point process on $(0,\infty)$ as represented by the K sequences of event-specific arrival times $\{t_i^k\}_{i\in\{1,\ldots,n^k\}}$, $k = 1,\ldots,K$, with counting functions $N^k(t) := \sum_{i\geq 1} \mathbb{1}_{\{t_i\leq t\}} \mathbb{1}_{\{W_i=k\}}$.

The internal history of an K-dimensional point process $N(t)$ is given by the filtration $\mathcal{F}_t^N$ with $\mathcal{F}_t^N = \sigma(N^k(s) : 0 \leq s \leq t, k \in \Xi)$, $N^k(s) = \sum_{i\geq 1} \mathbb{1}_{\{t_i\leq s\}} \mathbb{1}_{\{W_i\in\Xi\}}$, where Ξ denotes the σ-field of all subsets of $\{1,\ldots,K\}$. More general filtrations, including e.g. also processes of explanatory variables (covariates) $\{z_i\}_{i\in\{1,\ldots,n\}}$ are denoted by $\mathcal{F}_t$ with $\mathcal{F}_t^N \subseteq \mathcal{F}_t$.

Define $x_i := t_i - t_{i-1}$ with $i = 1,\ldots,n$ and $t_0 := 0$ as the inter-event duration from t_{i-1} until t_i. Furthermore, $x(t)$ with $x(t) := t - t_{\breve{N}(t)}$, with $\breve{N}(t) := \sum_{i\geq 1} \mathbb{1}_{\{t_i<t\}}$ denoting the left-continuous counting function, is called the backward recurrence time. It is a left-continuous function that grows linearly through time with discrete jumps back to zero *after* each arrival time t_i. Finally, let $\theta \in \Theta$ denote model parameters.

2.2 Compensators, intensities, and hazard rates

In martingale-based point process theory, the concept of *compensators* plays an important role. Using the property that an $\mathcal{F}_t$-adapted point process $N(t)$ is a submartingale[1], it can be decomposed into a zero mean martingale $M(t)$ and a (unique) $\mathcal{F}_t$-predictable increasing process, $\widetilde{\Lambda}(t)$, which is called the compensator of $N(t)$ and can be interpreted as the local conditional mean of $N(t)$ given the past. In statistical theory, this decomposition is typically referred to as the Doob-Meyer decomposition.

[1] An $\mathcal{F}_t$-adpated càdlàg process $N(t)$ is a submartingale if $\mathrm{E}[|N(t)|] < \infty$ for each t and if $s < t$ implies that $\mathrm{E}[N(t)|\mathcal{F}_s] \geq N(s)$.

Define $\lambda(t)$ as a scalar, positive $\mathcal{F}_t$-predictable process, i.e. $\lambda(t)$ is adapted to $\mathcal{F}_t$, and left-continuous with right hand limits. Then, $\lambda(t)$ is called the ($\mathcal{F}_t$-conditional) *intensity* of $N(t)$ if

$$\widetilde{\Lambda}(t) = \int_0^t \lambda(u)du, \tag{1}$$

where $\widetilde{\Lambda}(t)$ is the (unique) compensator of $N(t)$. This relationship emerges from the interpretation of the compensator as integrated (conditional) hazard function. Consequently, $\lambda(t)$ can be also defined by the relation

$$\mathrm{E}[N(s) - N(t)|\mathcal{F}_t] = \mathrm{E}\left[\int_t^s \lambda(u)du \Big| \mathcal{F}_t\right], \tag{2}$$

which has to hold (almost surely) for all t, s with $0 \leq t \leq s$. Letting $s \downarrow t$ leads to the heuristic representation which is more familiar in classical duration analysis. Then, $\lambda(t)$ is obtained by

$$\lambda(t+) := \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{E}\left[N(t+\Delta) - N(t) | \mathcal{F}_t\right], \tag{3}$$

where $\lambda(t+) := \lim_{\Delta \downarrow 0} \lambda(t+\Delta)$. In case of a stationary point process, $\bar{\lambda} := \mathrm{E}[dN(t)]/dt = \mathrm{E}[\lambda(t)]$ is constant.

Equation (3) manifests the close analogy between the intensity function and the *hazard function* which is given by

$$h(x) := f(x)/S(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr[x \leq X < x + \Delta | X \geq x] \tag{4}$$

with x denoting the (inter-event) duration as represented by the realization of a random variable X with probability density function $f(x)$, survivor function $S(x) := 1 - F(x)$, and cumulative distribution function (cdf) $F(x) := \Pr[X \leq x]$. Whereas the intensity function is defined in (continuous) calendar time, the hazard rate is typically defined in terms of the length of a duration x and is a key concept in (cross-section) survival analysis.

2.3 Types and representations of point processes

The simplest type of point process is the homogeneous Poisson process defined by

$$\Pr\left[(N(t+\Delta) - N(t)) = 1 \,|\mathcal{F}_t\right] = \lambda\Delta + o(\Delta), \tag{5}$$

$$\Pr\left[(N(t+\Delta) - N(t)) > 1 \,|\mathcal{F}_t\right] = o(\Delta), \tag{6}$$

with $\Delta \downarrow 0$. Then, $\lambda > 0$ is called the Poisson rate corresponding to the (constant) intensity. Accordingly, equations (5) and (6) define the *intensity representation* of a Poisson process. A well-known property of homogenous Poisson processes is that the inter-event waiting times x_i are independently exponentially distributed, leading to the *duration representation*. In this context, λ is the hazard rate of the exponential distribution. Furthermore, it can be shown (see e.g. Lancaster (1997)) that the number of events in an interval $(a, b]$, $N(a, b) := N(b) - N(a)$ is Poisson distributed with $\Pr[N(a, b) = k] = \exp[-\lambda(b - a)][\lambda(b - a)]^k/k!$, yielding the *counting representation*. All three representations of a Poisson process can be used as the starting point for the specification of a dynamic point process model.

Throughout this chapter we associate the term *duration models* to a model of the (discrete-time) duration process observable at the event-times $\{t_i\}_{i=1,\ldots,n}$. Then, researchers parameterize the conditional distribution function $F(x_i|\mathcal{F}_{t_{i-1}})$ or, alternatively, the conditional hazard rate $h(x_i|\mathcal{F}_{t_{i-1}})$. Generally, such a model should aim, in particular, at fitting the dynamical and distributional properties of durations. The latter is often characterized by the excess dispersion, corresponding to the ratio between the standard deviation to the mean. In classical hazard rate models employed in traditional survival analysis, the hazard rate is typically parameterized in terms of covariates, see e.g. Kalbfleisch and Prentice (1980), Kiefer (1988) or Lancaster (1997). The most well-known hazard model is the *proportional hazard* model introduced by Cox (1972) and is given by

$$h(x|z;\theta) = h_0(x|\gamma_1)g(z,\gamma_2), \tag{7}$$

where $\theta = (\gamma_1, \gamma_2)$, $h_0(\cdot)$ denotes the so-called baseline hazard rate and $g(\cdot)$ is a function of the covariates z and parameters γ_2. The baseline hazard rate may be parameterized in accordance with a certain distribution, like e.g., a Weibull distribution with parameters $\lambda, p > 0$ implying

$$h_0(x|\gamma_1) = \lambda p(\lambda x)^{p-1}. \tag{8}$$

For $p = 1$ we obtain the exponential case $h_0(x|\gamma_1) = \lambda$, implying a constant hazard rate. Alternatively, if $p > 1$, $\partial h_0(x|\gamma_1)/\partial x > 0$, i.e. the hazard rate is increasing with the length of the spell which is referred to as "positive duration dependence". In contrast, $p < 1$ implies "negative duration dependence". Non-monotonic hazard rates can be obtained with more flexible distributions, like the generalized F and special cases thereof, including the generalized gamma, Burr, Weibull and log-logistic distributions. We refer to the Appendix to Chapter 3 of Bauwens and Giot (2001) and to the Appendix of Hautsch (2004) for definitions and properties. Alternatively, the baseline hazard may be left unspecified and can be estimated nonparametrically, see Cox (1975).

An alternative type of duration model is the class of *accelerated failure time* (AFT) models given by

$$h(x|z;\theta) = h_0[xg(z,\gamma_2)|\gamma_1]g(z,\gamma_2). \tag{9}$$

Here, the effect of the exogenous variables is to accelerate or to decelerate the time scale on which the baseline hazard h_0 is defined. As illustrated in Section 3.1, AFT-type approaches are particularly attractive to model autocorrelated duration processes.

Because of their discrete-time nature, duration models cannot be used whenever the information set has to be updated *within* a duration spell, e.g. caused by time-varying covariates or event arrivals in other point processes. For this reason, (discrete-time) duration models are typically used in a univariate framework.

Whenever a continuous-time modelling is preferential (as e.g. to account for the asynchronous event arrivals in a multivariate framework), it is more natural to specify the intensity function directly, leading to so-called *intensity models*. One important extension of a homogenous Poisson process is to allow the intensity to be directed by a real-valued, non-negative (stationary) random process $\lambda^*(t)$ with (internal) history $\mathcal{F}_t^*$ leading to the class of *doubly stochastic Poisson processes* (Cox processes). In particular, $N(t)$ is called a Cox process directed by $\lambda^*(t)$ if conditional on $\lambda^*(t)$, $N(t)$ is a Poisson process with mean $\lambda^*(t)$, i.e. $\Pr[N(a,b) = k|\mathcal{F}_t^*] = \exp[-\lambda^*(t)]\,[\lambda^*(t)]^k\,/k!$. The doubly stochastic Poisson process yields a powerful class of probabilistic models with applications in seismology, biology and economics. For instance, specifying $\lambda^*(t)$ in terms of an autoregressive process yields a dynamic intensity model which is particularly useful to capture the clustering in financial point processes. For a special type of doubly stochastic Poisson process see Section 4.2.

A different generalization of the Poisson process is obtained by specifying $\lambda(t)$ as a (linear) self-exciting process given by

$$\lambda(t) = \omega + \int_0^t w(t-u)dN(u) = \omega + \sum_{t_i<t} w(t-t_i), \tag{10}$$

where ω is a constant, $w(s)$ denotes a non-negative weight function, and $\int_0^t w(s)dN(s)$ is the stochastic Stieltjes integral of the process w with respect to the counting process $N(t)$. The process (10) was proposed by Hawkes (1971) and is therefore named a *Hawkes* process. If $w(s)$ declines with s, then, the process is *self-exciting* in the sense that $\mathrm{Cov}[N(a,b), N(b,c)] > 0$, where $0 < a \leq b < c$. Different types of Hawkes processes and their applications to financial point processes are presented in Section 4.1. In alternative specifications, the intensity is driven by an autoregressive process which is updated at each point of the process. This leads to a special type of point process models which does not originate from the classical point process literature but rather from the *autoregressive conditional duration* (ACD) literature reviewed in Section 3 and brings time series analysis into play. The

resulting model is called an *autoregressive conditional intensity* model and is considered in Section 4.2.

Finally, starting from the counting representation of a Poisson process leads to the class of *count data models*. Dynamic extensions of Poisson processes in terms of counting representations are not surveyed in this chapter. Important references are e.g. Rydberg and Shephard (2003) and Heinen and Rengifo (2007).

2.4 The random time change theorem

One fundamental result of martingale-based point process theory is the (multivariate) random time change theorem by Meyer (1971) which allows to transform a wide class of point processes to a homogeneous Poisson process:

Theorem 1 *Assume that a multivariate point process* $(N^1(t), \ldots, N^K(t))$ *is formed from the event times* $\{t_i^k\}_{i\in\{1,\ldots,n^k\}}$, $k = 1, \ldots, K$, *and has continuous compensators* $(\widetilde{\Lambda}^1(t), \ldots, \widetilde{\Lambda}^K(t))$ *with* $\widetilde{\Lambda}^k(\infty) = \infty$ *for each* $k = 1, \ldots, K$. *Then, the point processes formed from* $\{\widetilde{\Lambda}^k(t_i^k)\}_{\{i=1,\ldots,n^k\}}$, $k = 1, \ldots, K$, *are independent Poisson processes with unit intensity.*

Proof. See Meyer (1971) or Brown and Nair (1988) for a more accessible and elegant proof. ∎

Define $\tau^k(t)$ as the $(\mathcal{F}_t$-)stopping time obtained by the solution of $\int_0^{\tau^k(t)} \lambda^k(s)ds = t$. Applying the random time change theorem to (1) implies that the point processes $\widetilde{N}^k(t)$ with $\widetilde{N}^k(t) := N^k(\tau^k(t))$ are independent Poisson processes with unit intensity and event times $\{\widetilde{\Lambda}^k(t_i^k)\}_{\{i=1,\ldots,n^k\}}$ for $k = 1, \ldots, K$. Then, the so-called integrated intensities

$$\Lambda^k(t_{i-1}^k, t_i^k) := \int_{t_{i-1}^k}^{t_i^k} \lambda^k(s)ds = \widetilde{\Lambda}^k(t_i^k) - \widetilde{\Lambda}^k(t_{i-1}^k) \tag{11}$$

correspond to the increments of independent Poisson processes for $k = 1, \ldots, K$. Consequently, they are independently standard exponentially distributed across i and k. For more details, see Bowsher (2007). The random time change theorem plays an important role in order to construct diagnostic tests for point process models (see Section 4.3) or to simulate point processes.

3 Dynamic Duration Models

In this section, we discuss univariate dynamic models for the durations between consecutive (financial) events. In Section 3.1, we review the class of ACD models, which is by far the most used class in the literature on financial point processes. In Section 3.2, we briefly discuss statistical inference for ACD models. In Section 3.3, we present other dynamic duration models, and in the last section we review some applications.

3.1 ACD models

The class of ACD models has been introduced by Engle and Russell (1997, 1998) and Engle (2000). In order to keep the notation simple, define x_i in the following as the inter-event duration which is standardized by a seasonality function $s(t_i)$, i.e. $x_i := (t_i - t_{i-1})/s(t_i)$. The function $s(t_i)$ is typically parameterized according to a spline function capturing time-of-day or day-of-week effects. Time-of-day effects arise because of systematic changes of the market activity throughout the day and due to openings of other related markets. In most approaches $s(t_i)$ is specified according to a linear or cubic spline function and is estimated separately in a first step yielding seasonality adjusted durations x_i. Alternatively, a non-parametric approach has been proposed by Veredas et al. (2002). For more details and examples regarding seasonality effects in financial duration processes, we refer the reader to Chapter 2 of Bauwens and Giot (2001) or to Chapter 3 of Hautsch (2004).

The key idea of the ACD model is to model the (seasonally adjusted) durations $\{x_i\}_{i=1,\ldots,n}$ in terms of a multiplicative error model in the spirit of Engle (2002), i.e.

$$x_i = \Psi_i \, \epsilon_i, \tag{12}$$

where Ψ_i denotes a function of the past durations (and possible covariates), and ε_i defines an i.i.d. random variable. It is assumed that

$$\mathrm{E}[\epsilon_\mathrm{i}] = 1, \tag{13}$$

so that Ψ_i corresponds to the conditional duration mean (the so-called "conditional duration") with $\Psi_i := \mathrm{E}[\mathrm{x_i}|\mathcal{F}_{\mathrm{t_{i-1}}}]$. The ACD model can be rewritten in terms of the intensity function as

$$\lambda(t|\mathcal{F}_t) = \lambda_\epsilon \left(\frac{x(t)}{\Psi_{\breve{N}(t)+1}} \right) \frac{1}{\Psi_{\breve{N}(t)+1}}, \tag{14}$$

where $\lambda_\epsilon(s)$ denotes the hazard function of the ACD error term. This formulation shows that the ACD model belongs to the class of AFT models.

Assuming ϵ_i to be standard exponentially distributed yields the so-called *Exponential ACD* model. More flexible specifications arise by assuming ϵ_i to follow a more general distribution, see the discusssion after equation (8). It is evident that the ACD model is the counter-part to the GARCH model (Bollerslev (1986)) for duration processes. Not surprisingly, many results and specifications from the GARCH literature have been adapted to the ACD literature.

The conditional duration, Ψ_i, is defined as a function Ψ of the information set $\mathcal{F}_{t_{i-1}}$ and provides therefore the vehicle for incorporating the dynamics of the duration process. In this respect it is convenient to use an ARMA-type structure of order (p, q), whereby

$$\Psi_i = \Psi(\Psi_{i-1}, \ldots, \Psi_{i-q}, x_{i-1}, \ldots, x_{i-p}). \tag{15}$$

For simplicity, we limit the exposition in the sequel to the case $p = q = 1$.

The first model put forward in the literature is the *linear ACD model*, which specializes (15) as

$$\Psi_i = \omega + \beta\Psi_{i-1} + \alpha x_{i-1}. \tag{16}$$

Since Ψ_i must be positive, the restrictions $\omega > 0$, $\alpha \geq 0$ and $\beta \geq 0$ are usually imposed. It is also assumed that $\beta = 0$ if $\alpha = 0$, otherwise β is not identified. The process defined by (12), (13) and (16) is covariance-stationary if

$$(\alpha + \beta)^2 - \alpha^2\sigma^2 < 1, \tag{17}$$

where $\sigma^2 := \mathrm{Var}[\epsilon_i] < \infty$, and has the following moments and autocorrelations:

(1) $\mathrm{E}[x_i] := \mu_x = \omega/(1 - \alpha - \beta)$,
(2) $\mathrm{Var}[x_i] := \sigma_x^2 = \mu_x^2\,\sigma^2 \frac{1-\beta^2-2\alpha\beta}{1-(\alpha+\beta)^2-\alpha^2\sigma^2}$,
(3) $\rho_1 = \frac{\alpha\,(1-\beta^2-\alpha\,\beta)}{1-\beta^2-2\,\alpha\,\beta}$ and $\rho_n = (\alpha + \beta)\rho_{n-1}$ for $n \geq 2$.

The condition (17) ensures the existence of the variance. These results are akin to those for the GARCH(1,1) zero-mean process. They can be generalized to ACD(p,q) processes with $p, q > 1$. In applications, estimates of $\alpha + \beta$ are typically found to be in the interval (0.85,1) with α lying in the interval (0.01,0.15). Since the ACD(1,1) model can be written as

$$x_i = \omega + (\alpha + \beta)x_{i-1} + u_i - \beta u_{i-1}, \tag{18}$$

where $u_i := x_i - \Psi_i$ is a martingale difference innovation, the resulting autocorrelation function (ACF) is that of an ARMA(1,1) process that has AR and MA roots close to each other. This type of parameter configuration generates the typical ACF shape of clustered data. Nevertheless, the ACF decreases at a geometric rate, though it is not uncommon to find duration series with an ACF that decreases at a hyperbolic rate. This tends to happen for long

series and may be due to instabilities of parameters which give the illusion of long memory in the process. In order to allow for long range dependence in financial duration processes, Jasiak (1998) extends the ACD model to a *fractionally integrated ACD* model. For alternative ways to specify long memory ACD models, see Koulikov (2002).

A drawback of the linear ACD model is that it is difficult to allow Ψ_i to depend on functions of covariates without violating the non-negativity restriction. For this reason, Bauwens and Giot (2000) propose a class of *logarithmic ACD models*, where no parametric restrictions are needed to ensure positiveness of the process:

$$\ln \Psi_i = \omega + \beta \ln \Psi_{i-1} + \alpha g(\epsilon_{i-1}), \tag{19}$$

where $g(\epsilon_{i-1})$ is either $\ln \epsilon_{i-1}$ (log-ACD of type I) or ϵ_{i-1} (type II). Using this setting, it is convenient to augment Ψ_i by functions of covariates, see e.g. Bauwens and Giot (2000). The stochastic process defined by (12), (13) and (19) is covariance-stationary if

$$\beta < 1, \quad \mathrm{E}\left[\epsilon_i \exp\{\alpha g(\epsilon_i)\}\right] < \infty, \quad \mathrm{E}\left[\exp\{2\alpha g(\epsilon_i)\}\right] < \infty. \tag{20}$$

Its mean, variance and autocorrelations are given in Section 3.2 in Bauwens and Giot (2001), see also Fernandes and Grammig (2006) and Bauwens et al. (2008). Drost and Werker (2004) propose to combine one of the previous ACD equations for the conditional duration mean with an unspecified distribution for ϵ_i, yielding a class of *semi-parametric ACD models*.

The *augmented ACD* (AACD) model introduced by Fernandes and Grammig (2006) provides a more flexible specification of the conditional duration equation than the previous models. Here, Ψ_i is specified in terms of a power transformation yielding

$$\Psi_i^{\delta_1} = \omega + \beta \Psi_{i-1}^{\delta_1} + \alpha \Psi_{i-1}^{\delta_1} [|\epsilon_{i-1} - \xi| - \rho(\epsilon_{i-1} - \xi)]^{\delta_2},$$

where $\delta_1 > 0$, $\delta_2 > 0$, ξ, and ρ are parameters. The so-called news impact function $[|\epsilon_{i-1} - \xi| - \rho(\epsilon_{i-1} - \xi)]^{\delta_2}$ allows a wide variety of shapes of the curve tracing the impact of ϵ_{i-1} on Ψ_i for a given value of Ψ_{i-1} and the remaining parameters. The parameters ξ and ρ are shift and rotation parameters, respectively. If $\xi = \rho = 0$, the linear ACD model is obtained by setting $\delta_1 = \delta_2 = 1$, the type I logarithmic ACD model by letting δ_1 and δ_2 tend to 0, and the type II version by letting δ_1 tend to 0 and setting $\delta_2 = 1$. Fernandes and Grammig (2006) compare different versions of the AACD model using IBM price durations arising from trading at the New York Stock Exchange (NYSE). Their main finding is that "letting δ_1 free to vary and accounting for asymmetric effects (by letting ξ and ρ free) seem to operate as substitute sources of flexibility". Hautsch (2006) proposes an even more general augmented ACD model that nests in particular the so-called *EXponential ACD model* proposed by Dufour and Engle (2000) implying a kinked news impact

function. As a counterpart to the so-called semiparametric GARCH model proposed by Engle and Ng (1993), Hautsch (2006) suggests specifying the news impact function in terms of a linear spline function based on the support of ε_i. He illustrates that the high flexibility of this model is needed in order to appropriately capture the dynamic properties of financial durations.

Another way to achieve flexibility in ACD models is to use the idea of mixtures. The mixture may apply to the error distribution alone, as in De Luca and Zuccolotto (2003), De Luca and Gallo (2004) and Hujer and Vuletic (2007), or may involve the dynamic component as well. Zhang et al. (2001) propose a *threshold ACD model* (TACD), wherein the ACD equation and the error distribution change according to a threshold variable such as the previous duration. For J regimes indexed by $j = 1, \ldots, J$, the model is defined as

$$x_i = \Psi_i \, \epsilon_i^{(j)}, \tag{21}$$

$$\Psi_i = \omega^{(j)} + \beta^{(j)} \Psi_{i-1} + \alpha^{(j)} x_{i-1} \tag{22}$$

when $x_{i-1} \in [r_{j-1}, r_j)$, and $0 = r_0 < r_1 < \ldots < r_J = \infty$ are the threshold parameters. The superscript (j) indicates that the distribution or the model parameters can vary with the regime operating at observation i. This model can be viewed as a mixture of J ACD models, where the probability to be in regime j at i is equal to 1 and the probabilities to be in each of the other regimes are equal to 0. Hujer et al. (2002) extend this model to let the regime changes be governed by a hidden Markov chain.

While the TACD model implies discrete transitions between the individual regimes, Meitz and Teräsvirta (2006) propose a class of *smooth transition ACD (STACD) models* generalizing linear and logarithmic ACD models. Conditions for strict stationarity, ergodicity, and existence of moments for this model and other ACD models are provided in Meitz and Saikkonen (2004) using the theory of Markov chains. A motivation for the STACD model is, like for the AACD model, to allow for a nonlinear impact of the past duration on the next expected duration.

3.2 Statistical inference

The estimation of most ACD models can be easily performed by maximum likelihood (ML). Engle (2000) demonstrates that the results by Bollerslev and Wooldridge (1992) on the quasi-maximum likelihood (QML) property of the Gaussian GARCH(1,1) model extend to the Exponential-ACD(1,1) model. QML estimates are obtained by maximizing the quasi-loglikelihood function given by

$$\ln \mathcal{L}\left(\theta; \{x_i\}_{\{i=1,\ldots,n\}}\right) = -\sum_{i=1}^{n}\left[\ln \Psi_i + \frac{x_i}{\Psi_i}\right]. \tag{23}$$

For more details we refer to Chapter 3 of Bauwens and Giot (2001), Chapter 5 of Hautsch (2004), and to the survey of Engle and Russell (2005).

Residual diagnostics and goodness-of-fit tests can be performed by evaluating the stochastic properties of the ACD residuals $\widehat{\epsilon}_i = x_i/\widehat{\Psi}_i$. The dynamic properties are easily analyzed based on Portmanteau statistics or tests against independence such as proposed by Brock et al. (1996). The distributional properties can be evaluated by Engle and Russell's (1998) test for no excess dispersion using the asymptotically standard normal test statistic $\sqrt{n/8}\,\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ denotes the empirical variance of the residual series. Dufour and Engle (2000) and Bauwens et al. (2004) evaluate the models' goodness-of-fit based on density forecasts using the probability integral transform as proposed by Diebold et al. (1998). A nonparametric test against distributional misspecification is proposed by Fernandes and Grammig (2005) based on the work by Aït-Sahalia (1996). Statistics that exclusively test for misspecifications of the conditional mean function Ψ_i have been worked out by Meitz and Teräsvirta (2006) using the Lagrange Multiplier principle and by Hautsch (2006) using (integrated) conditional moment tests. A common result is that too simple ACD specifications, such as the ACD or Log-ACD model are not flexible enough to adequately capture, even in-sample, the properties of observed financial durations. However, in order to avoid the problem of potential over-fitting a serious comparison of ACD specifications should rely on out-of-sample evaluations.

3.3 Other models

ACD models strongly resemble ARCH models. Therefore it is not surprising that Taylor's (1986) stochastic volatility model for financial returns has been a source of inspiration of corresponding duration models. Bauwens and Veredas (2004) propose the *stochastic conditional duration (SCD) model* as an alternative to ACD-type models. The SCD model relates to the logarithmic ACD model in the same way as the stochastic volatility model relates to (a restricted version of) the exponential GARCH model by Nelson (1991). Thus the model is defined by equations (12), (13), and

$$\ln \Psi_i = \omega + \beta \ln \Psi_{i-1} + \gamma \epsilon_{i-1} + u_i, \tag{24}$$

where u_i is iid $\mathrm{N}(0, \sigma_u^2)$. The process $\{u_i\}$ is assumed to be independent of the process $\{\epsilon_i\}$. The set of possible distributions for the duration innovations ϵ_i is the same as that for ACD models. This model generates a rich class of hazard functions for x_i through the interplay of two distributions. The latent variable

Ψ_i may be interpreted as being inversely related to the information arrival process which triggers bursts of activity on financial markets. The "leverage" term $\gamma\epsilon_{i-1}$ in (24) is added by Feng et al. (2004) to allow for an intertemporal correlation between the observable duration and the conditional duration. Bauwens and Veredas (2004) use a logarithmic transformation of (12) and employ QML estimation based on the Kalman filter. Strickland et al. (2006) use Bayesian estimation with a Markov chain Monte Carlo algorithm. For ML estimation, Feng et al. (2004) use the Monte Carlo method of Durbin and Koopman (2004), and Bauwens and Galli (2008) use efficient importance sampling.

The ACD and SCD models reviewed above share the property that the dynamics of higher moments of the duration process are governed by the dynamics of the conditional mean. Ghysels et al. (2004) argue that this feature is restrictive and introduce a nonlinear two factor model that disentangles the movements of the mean and that of the variance of durations. Since the second factor is responsible for the variance heterogeneity, the model is named the *stochastic volatility duration (SVD) model*. The departure point for this model is a standard static duration model in which the durations are independently and exponentially distributed with a gamma heterogeneity, i.e.

$$x_i = \frac{U_i}{aV_i} = \frac{H(1, F_{1i})}{aH(b, F_{2i})}, \tag{25}$$

where U_i and V_i are two independent variables which are gamma(1,1) (i.e. exponential) and gamma(b, b) distributed, respectively. The last ratio in (25) uses two independent Gaussian factors F_{1i} and F_{2i}, and $H(b, F) = G(b, \varphi(F))$, where $G(b, .)$ is the quantile function of the gamma(b, b) distribution and $\varphi(.)$ the cdf of the standard normal distribution. Ghysels et al. (2004) extend this model to a dynamic setup through a VAR model for the two underlying Gaussian factors. The estimation of the model requires simulation methods.

3.4 Applications

ACD models can be used to estimate and predict the intra-day volatility of returns from the intensity of price durations. As shown by Engle and Russell (1998), a price intensity is closely linked to the instantaneous price change volatility. The latter is given by

$$\widetilde{\sigma}^2(t) := \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{E}\left[\left(\frac{p(t+\Delta) - p(t)}{p(t)} \right)^2 \middle| \mathcal{F}_t \right], \tag{26}$$

where $p(t)$ denotes the price (or midquote) at t. By denoting the counting process associated with the event times of cumulated absolute price changes of size dp by $N^{dp}(t)$, we can formulate (26) in terms of the intensity function of the process of dp-price changes. Then, the dp-price change instantaneous volatility can be computed as

$$\begin{aligned}\widetilde{\sigma}^2_{(dp)}(t) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr\left[|p(t+\Delta) - p(t)| \geq dp \,|\mathcal{F}_t\right] \cdot \left[\frac{dp}{p(t)}\right]^2 \\ &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr\left[\{N^{dp}(t+\Delta) - N^{dp}(t)\} > 0 \,|\mathcal{F}_t\right] \cdot \left[\frac{dp}{p(t)}\right]^2 \\ &:= \lambda^{dp}(t) \cdot \left[\frac{dp}{p(t)}\right]^2, \end{aligned} \tag{27}$$

where $\lambda^{dp}(t)$ denotes the corresponding dp-price change intensity. Hence, using (14), one can estimate or predict the instantaneous volatility of the price process $p(t)$ at any time point. Giot (2005) compares these estimates with usual GARCH based estimates obtained by interpolating the prices on a grid of regularly spaced time points. He finds that GARCH based predictions are better measures of risk than ACD based ones in a Value-at-Risk (VaR) evaluation study.

ACD and related models have been typically used to test implications of asymmetric information models of price formation. For example, the model of Easley and O'Hara (1992) implies that the number of transactions influences the price process through information based clustering of transactions. Then, including lags as well as expectations of the trading intensity as explanatory variables in a model for the price process allows to test such theoretical predictions. For a variety of different applications in market microstructure research, see Engle and Russell (1998), Engle (2000), Bauwens and Giot (2000), Engle and Lunde (2003), and Hafner (2005) among others. Several authors have combined an ACD model with a model for the marks of a financial point process. The basic idea is to model the duration process by an ACD model, and conditionally on the durations, to model the process of marks. Bauwens and Giot (2003) model the direction of the price change between two consecutive trades by formulating a competing risks model, where the direction of the price movement is triggered by a Bernoulli process. Then, the parameters of the ACD process depend on the direction of the previous price change, leading to an asymmetric ACD model. A related type of competing risks model is specified by Bisière and Kamionka (2000). Prigent et al. (2001) use a similar model for option pricing. Russell and Engle (2005) develop an autoregressive conditional multinomial model to simultaneously model the time between trades and the dynamic evolution of (discrete) price changes. An alternative approach based on a dynamic integer count model is proposed by Liesenfeld et al. (2006).

A related strand of the literature studies the interaction between the trading intensity and the trade-to-trade return volatility. Engle (2000) augments a GARCH equation for returns per time unit by the impact of the inverse of the observed and expected durations (x_i and Ψ_i), and of the surprise x_i/Ψ_i. A decrease in x_i or Ψ_i has a positive impact on volatility while the surprise has a reverse influence. Ghysels and Jasiak (1998) and Grammig and Wellner (2002) study a GARCH process for trade-to-trade returns with time-varying parameters which are triggered by the trading intensity. Meddahi et al. (2006) derive a discrete time GARCH model for irregularly spaced data from a continuous time volatility process and compare it to the ACD-GARCH models by Engle (2000) and Ghysels and Jasiak (1998).

4 Dynamic Intensity Models

In this section, we review the most important types of dynamic intensity models which are applied to model financial point processes. The class of Hawkes processes and extensions thereof are discussed in Section 4.1. In Section 4.2, we survey different autoregressive intensity models. Statistical inference for intensity models is presented in Section 4.3, whereas the most important applications in the recent literature are briefly discussed in Section 4.4.

4.1 Hawkes processes

Hawkes processes originate from the statistical literature in seismology and are used to model the occurrence of earthquakes, see e.g. Vere-Jones (1970), Vere-Jones and Ozaki (1982), and Ogata (1988) among others. Bowsher (2007) was the first applying Hawkes models to financial point processes. As explained in Section 3.2, Hawkes processes belong to the class of self-exciting processes, where the intensity is driven by a weighted function of the time distance to previous points of the process. A general class of univariate Hawkes processes is given by

$$\lambda(t) = \varphi\left(\mu(t) + \textstyle\sum_{t_i<t} w(t-t_i)\right), \tag{28}$$

where φ denotes a possibly nonlinear function, $\mu(t)$ is a deterministic function of time, and $w(s)$ denotes a weight function. If φ is a positive function, we obtain the class of nonlinear Hawkes processes considered by Brémaud and Massoulié (1996). In this case, $\mu(t)$ and $w(t)$ can take negative values since the transformation $\varphi(\cdot)$ preserves the non-negativity of the process. Such a specification is useful whenever the intensity may be negatively affected by the process history or covariates. For instance, in the context of financial

duration processes, $\mu(t)$ can be parameterized as a function of covariates. Stability conditions for nonlinear Hawkes processes are derived by Brémaud and Massoulié (1996). For the special case where φ is a linear function, we obtain the class of linear Hawkes processes originally considered by Hawkes (1971). They are analytically and computationally more tractable than their nonlinear counterparts, however, they require $\mu(t) > 0$ and $w(t) > 0$ in order to ensure non-negativity.

As pointed out by Hawkes and Oakes (1974), linear self-exciting processes can be viewed as clusters of Poisson processes. Then, each event is one of two types: an immigrant process or an offspring process. The immigrants follow a Poisson process and define the centers of so-called Poisson clusters. If we condition on the arrival time, say t_i, of an immigrant, then independently of the previous history, t_i is the center of a Poisson process, $\Upsilon(t_i)$, of offspring on (t_i, ∞) with intensity function $\lambda_i(t) = \lambda(t - t_i)$, where λ is a non-negative function. The process $\Upsilon(t_i)$ defines the first generation offspring process with respect to t_i. Furthermore, if we condition on the process $\Upsilon(t_i)$, then each of the events in $\Upsilon(t_i)$, say t_j, generates a Poisson process with intensity $\lambda_j(t) = \lambda(t - t_j)$. These independent Poisson processes build the second generation of offspring with respect to t_i. Similarly, further generations arise. The set of all offspring points arising from one immigrant are called a Poisson cluster. Exploiting the branching and conditional independence structure of a (linear) Hawkes process, Møller and Rasmussen (2004) develop a simulation algorithm as an alternative to the Shedler-Lewis thinning algorithm or the modified thinning algorithm by Ogata (1981) (see e.g. Daley and Vere-Jones (2003)). The immigrants and offsprings can be referred to as "main shocks" and "after shocks" respectively. This admits an interesting interpretation which is useful not only in seismology but also in high-frequency finance. Bowsher (2007), Hautsch (2004) and Large (2007) illustrate that Hawkes processes capture the dynamics of financial point processes remarkably well. This indicates that the cluster structure implied by the self-exciting nature of Hawkes processes seem to be a reasonable description of the timing structure of events on financial markets.

The most common parameterization of $w(t)$ has been suggested by Hawkes (1971) and is given by

$$w(t) = \sum_{j=1}^{P} \alpha_j e^{-\beta_j t}, \tag{29}$$

where $\alpha_j \geq 0$, $\beta_j > 0$ for $j = 1, \ldots, P$ are model parameters, and P denotes the order of the process and is selected exogenously (or by means of information criteria). The parameters α_j are scale parameters, whereas β_j drive the strength of the time decay. For $P > 1$, the intensity is driven by the superposition of different exponentially decaying weighted sums of the backward times to all previous points. In order to ensure identification we impose the constraint $\beta_1 > \ldots > \beta_P$. It can be shown that the stationarity of the process

requires $0 < \int_0^\infty w(s)ds < 1$, which is ensured only for $\sum_{j=1}^P \alpha_j/\beta_j < 1$, see Hawkes (1971).

While (29) implies an exponential decay, the alternative parameterization

$$w(t) = \frac{H}{(t+\kappa)^p}, \tag{30}$$

with parameters H, κ, and $p > 1$ features a hyperbolic decay. Such weight functions are typically applied in seismology and allow to capture long range dependence. Since financial duration processes also tend to reveal long memory behavior (see Jasiak (1998)), this specification may be interesting in financial applications.

Multivariate Hawkes models are obtained by a generalization of (28). Then, $\lambda(t)$ is given by the $(K \times 1)$-vector $\lambda(t) = (\lambda^1(t), \ldots, \lambda^K(t))'$ with

$$\lambda^k(t) = \varphi\left(\mu^k(t) + \textstyle\sum_{r=1}^K \sum_{t_i^r < t} w_r^k(t - t_i^r)\right), \tag{31}$$

where $w_r^k(s)$ is a k-type weight function of the backward time to all r-type events. Using an exponential decay function, Hawkes (1971) suggests to parameterize $w_r^k(s)$ as

$$w_r^k(t) = \sum_{j=1}^P \alpha_{r,j}^k e^{-\beta_{r,j}^k t}, \tag{32}$$

where $\alpha_{r,j}^k \geq 0$ and $\beta_{r,1}^k > \ldots > \beta_{r,P}^k > 0$ determine the influence of the time distance to past r-type events on the k-type intensity. Thus, in the multivariate case, $\lambda^k(t)$ depends not only on the distance to all k-type points, but also on the distance to all other points of the pooled process. Hawkes (1971) provides a set of linear parameter restrictions ensuring the stationarity of the process.

Bowsher (2007) proposes a generalization of the Hawkes model which allows to model point processes that are interrupted by time periods where no activity takes place. In high-frequency financial time series these effects occur because of trading breaks due to trading halts, nights, weekends or holidays. In order to account for such effects, Bowsher proposes to remove all non-activity periods and to concatenate consecutive activity periods by a spill-over function.

4.2 Autoregressive intensity processes

Hamilton and Jordà (2002) establish a natural link between ACD models and intensity models by allowing the ACD model to include covariates that may change during a duration spell (time-varying covariates). Their so-called

autoregressive conditional hazard (ACH) model relies on the idea that in the Exponential ACD model, the intensity corresponds to the inverse of the conditional duration, i.e. $\lambda(t) = \Psi^{-1}_{\check{N}(t)+1}$. They extend this expression by a function of time-varying regressors $z_{\widetilde{t}_j}$, where $\widetilde{t}_j$ denotes the arrival times in the covariate process and j is such that $t_{\check{N}(t)} < \widetilde{t}_j < t$. Then,

$$\lambda(t) = \frac{1}{\Psi_{\check{N}(t)+1} + z'_{\widetilde{t}_j}\gamma}, \tag{33}$$

where γ is a vector of unknown parameters.

An alternative model which can be seen as a combination of a duration model and an intensity model is introduced by Gerhard and Hautsch (2007). They propose a dynamic extension of a proportional intensity model due to Cox (1972), where the baseline intensity $\lambda_0(t)$ is not specified. Their key idea is to exploit the stochastic properties of the integrated intensity and to re-formulate the model in terms of a regression model with unknown left-hand variable and Gumbel distributed error terms. See Kiefer (1988) for a nice illustration of this relation. To identify the unknown baseline intensity at discrete points, Gerhard and Hautsch follow the idea of Han and Hausman (1990) and formulate the model in terms of an ordered response model based on categorized durations. In order to allow for serial dependence in the duration process, the model is extended by an observation-driven ARMA structure based on generalized errors. As a result, the resulting *semiparametric autoregressive conditional proportional intensity model* allows to capture serial dependence in durations and to estimate conditional failure probabilities without requiring explicit distributional assumptions.

In the *autoregressive conditional intensity (ACI) models* introduced by Russell (1999), the intensity function is directly modeled as an autoregressive process which is updated by past realizations of the integrated intensity. Let $\lambda(t) = (\lambda^1(t), \ldots, \lambda^K(t))'$. Russell (1999) proposes to specify $\lambda^k(t)$ in terms of a proportional intensity structure given by

$$\lambda^k(t) = \Phi^k_{\check{N}(t)+1}\lambda^k_0(t)s^k(t), \qquad k = 1, \ldots K, \tag{34}$$

where $\Phi_{\check{N}(t)+1}$ captures the dynamic structure, $\lambda^k_0(t)$ is a baseline intensity component capturing the (deterministic) evolution of the intensity between two consecutive points and $s^k(t)$ denotes a deterministic function of t capturing, for instance, possible seasonality effects. The function $\Phi_{\check{N}(t)}$ is indexed by the left-continuous counting function and is updated instantaneously *after* the arrival of a new point. Hence, Φ_i is constant for $t_{i-1} < t \leq t_i$. Then, the evolution of the intensity function between two consecutive arrival times is governed by $\lambda^k_0(t)$ and $s^k(t)$.

In order to ensure the non-negativity of the process, the dynamic component Φ^k_i is specified in log-linear form, i.e.

$$\Phi_i^k = \exp\left(\widetilde{\Phi}_i^k + z_{i-1}'\gamma^k\right), \tag{35}$$

where z_i denotes a vector of explanatory variables observed at arrival time t_i and γ^k is the corresponding parameter vector. Define ε_i as a scalar innovation term which is computed from the integrated intensity function associated with the most recently observed process, i.e.

$$\varepsilon_i := \sum_{k=1}^{K}\left(1 - \int_{t^k_{N^k(t_i)-1}}^{t^k_{N^k(t_i)}} \lambda^k(s)ds\right) y_i^k, \tag{36}$$

where y_i^k defines an indicator variable that takes on the value one if the i-th point of the pooled process is of type k and zero otherwise. According to the random time change argument presented in Section 2.4, ε_i corresponds to a random mixture of i.i.d. centered standard exponential variates and thus is itself an i.i.d. zero mean random variable. Then, the $(K \times 1)$ vector $\widetilde{\Phi}_i = \left(\widetilde{\Phi}_i^1, \ldots, \widetilde{\Phi}_i^K\right)'$ is parameterized as

$$\widetilde{\Phi}_i = \sum_{k=1}^{K}\left(A^k\varepsilon_{i-1} + B^k\widetilde{\Phi}_{i-1}\right) y_{i-1}^k, \tag{37}$$

where $A^k = \{a_j^k\}$ denotes a $(K \times 1)$ parameter vector and $B^k = \{b_{ij}^k\}$ is a $(K \times K)$ matrix of persistence parameters. Hence, the fundamental principle of the ACI model is that at each event t_i all K processes are updated by the realization of the integrated intensity with respect to the most recent process, where the impact of the innovation on the K processes can be different and also varies with the type of the most recent point. As suggested by Bowsher (2007), an alternative specification of the ACI innovation term could be $\widetilde{\varepsilon}_i = 1 - \Lambda(t_{i-1}, t_i)$, where $\Lambda(t_{i-1}, t_i) := \sum_{k=1}^{K}\Lambda^k(t_{i-1}, t_i)$ denotes the integrated intensity of the pooled process computed between the two most recent points. Then, following the arguments above, $\widetilde{\varepsilon}_i$ is a zero mean i.i.d. innovation term. Because of the regime-switching nature of the persistence matrix, the derivation of stationarity conditions is difficult. However, a sufficient (but not necessary) condition is that the eigenvalues of the matrices B^k for all $k = 1, \ldots, K$ lie inside the unit circle.

As proposed by Hautsch (2004), the baseline intensity function $\lambda_0^k(t)$ can be specified as the product of K different Burr hazard rates, i.e.

$$\lambda_0^k(t) = \exp(\omega^k)\prod_{r=1}^{K}\frac{x^r(t)^{p_r^s-1}}{1+\eta_r^s x^r(t)^{p_r^s}}, \quad p_r^s > 0,\ \eta_r^s \geq 0. \tag{38}$$

According to this specification $\lambda^k(t)$ is driven not only by the k-type backward recurrence time but also by the time distance to the most recent point in all other processes $r = 1, \ldots, K$ with $r \neq k$. A special case occurs when $p_r^s = 1$

and $\eta_r^s = 0$, $\forall\ r \neq s$. Then, the k-th process is affected only by its own history.

Finally, $s^k(t)$ is typically specified as a spline function in order to capture intraday seasonalities. A simple parameterization which is used in most studies is given by a linear spline function of the form $s^k(t) = 1 + \sum_{j=1}^{S} \nu_j^k (t - \tau_j) \cdot \mathbb{1}_{\{t>\tau_j\}}$, where τ_j, $j = 1 \ldots, S$, denote S nodes within a trading period and ν_j the corresponding parameters. A more flexible parameterization is e.g. given by a flexible Fourier form (Gallant (1981)) as used by Andersen and Bollerslev (1998) or Gerhard and Hautsch (2002) among others.

If $K = 1$ and $\eta_1^1 = 0$, the ACI model corresponds to a re-parameterized form of the Log-ACD model. If the ACI model is extended to include time-varying covariates (see Hall and Hautsch (2007)), it generalizes the approach by Hamilton and Jordà (2002). In this case, all event times associated with (discrete time) changes of time-varying covariates are treated as another point process that is not explicitly modelled. Then, at each event time of the covariate process, the multivariate intensity is updated, which requires a piecewise computation of the corresponding integrated intensities.

A generalization of the ACI model has been proposed by Bauwens and Hautsch (2006). The key idea is that the multivariate intensity function $\lambda(t) = (\lambda^1(t), \ldots, \lambda^K(t))'$ is driven not only by the observable history of the process but also by a common component. The latter may be considered as a way to capture the unobservable general information flow in a financial market. By assuming the existence of a common unobservable factor $\lambda^*(t)$ following a pre-assigned structure in the spirit of a doubly stochastic Poisson process (see Section 2.3), we define the internal (unobservable) history of $\lambda^*(t)$ as $\mathcal{F}_t^*$. We assume that $\lambda(t)$ is adapted to the filtration $\mathcal{F}_t := \sigma(\mathcal{F}_t^o \cup \mathcal{F}_t^*)$, where $\mathcal{F}_t^o$ denotes some *observable* filtration. Then, the so-called *stochastic conditional intensity (SCI) model* is given by

$$\lambda^k(t) = \lambda^{o,k}(t) \left(\lambda^*_{\breve{N}(t)+1}\right)^{\sigma_k^*}, \tag{39}$$

where $\lambda^*_{\breve{N}(t)+1} := \lambda^*(t_{\breve{N}(t)+1})$ denotes the common latent component which is updated at each point of the (pooled) process $\{t_i\}_{i\in\{1,\ldots,n\}}$. The direction and magnitude of the process-specific impact of λ^* is driven by the parameters σ_k^*. The process-specific function $\lambda^{o,k}(t)$ denotes a conditionally deterministic idiosyncratic k-type intensity component given the *observable* history, $\mathcal{F}_t^o$.

Bauwens and Hautsch (2006) assume that λ_i^* has left-continuous sample paths with right-hand limits and follows a log-linear zero mean AR(1) process given by

$$\ln \lambda_i^* = a^* \ln \lambda_{i-1}^* + u_i^*, \qquad u_i^* \sim \text{iid } N(0,1). \tag{40}$$

Because of the symmetry of the distribution of $\ln \lambda_i^*$, Bauwens and Hautsch impose an identification assumption which restricts the sign of one of the scaling parameters σ_k^*. The observation-driven component $\lambda^{o,k}(t)$ is specified

as the ACI model described above. However, in contrast to the basic ACI model, in the SCI model, the innovation term is computed based on the *observable* history of the process, i.e.

$$\varepsilon_i = \sum_{k=1}^{K} \left\{ -\varpi - \ln \Lambda^{o,k} \left(t^k_{N^k(t_i)-1}, t^k_{N^k(t_i)} \right) \right\} y^k_i, \tag{41}$$

where ϖ denotes Euler's constant, $\varpi = 0.5772$. Here, $\Lambda^{o,k}\left(t^k_{i-1}, t^k_i\right)$ is given by

$$\begin{aligned} \Lambda^{o,k}\left(t^k_{i-1}, t^k_i\right) &:= \sum_{j=N(t^k_{i-1})}^{N(t^k_i)-1} \int_{t_j}^{t_{j+1}} \lambda^{o,k}(u)du \\ &= \sum_{j=N(t^k_{i-1})}^{N(t^k_i)-1} \left(\lambda^*_j\right)^{-\sigma^*_k} \Lambda^k\left(t_j, t_{j+1}\right) \end{aligned} \tag{42}$$

corresponding to the sum of (piecewise) integrated k-type intensities which are observed through the duration spell and are standardized by the corresponding (scaled) realizations of the latent component. This specification ensures that ε_i can be computed exclusively based on past observables implying a separation between the observation-driven and the parameter-driven components of the model. Bauwens and Hautsch (2006) analyze the probabilistic properties of the model and illustrate that the SCI model generates a wide range of (cross-)autocorrelation structures in multivariate point processes. In an application to a multivariate process of price intensities, they find that the latent component captures a substantial part of the cross-dependencies between the individual processes resulting in a quite parsimonious model. An extension of the SCI model to the case of multiple states is proposed by Koopman et al. (2008) and is applied to the modelling of credit rating transitions.

4.3 Statistical inference

Karr shows that valid statistical inference can be performed based on the intensity function solely, see Theorem 5.2. in Karr (1991) or Bowsher (2007). Assume a K-variate point process $N(t) = \{N^k(t)\}_{k=1}^K$ on $(0, T]$ with $0 < T < \infty$, and the existence of a K-variate $\mathcal{F}_t$-predictable process $\lambda(t)$ that depends on the parameters θ. Then, it can be shown that a genuine log likelihood function is given by

$$\ln \mathcal{L}\left(\theta; \{N(t)\}_{t\in(0,T]}\right) = \sum_{k=1}^{K}\left[\int_0^T (1-\lambda^k(s))ds + \int_{(0,T]} \ln \lambda^k(s) dN^k(s)\right],$$

which can be alternatively computed by

$$\ln \mathcal{L}\left(\theta; \{N(t)\}_{t\in(0,T]}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K}(-\Lambda^k(t_{i-1},t_i)) + y_i^k \ln\left[\lambda^k(t_i)\right] + TK. \quad (43)$$

Note that (43) differs from the standard log likelihood function of duration models by the additive (integrating) constant TK which can be ignored for ML estimation. By applying the so-called exponential formula (Yashin and Arjas (1988)), the relation between the integrated intensity function and the conditional survivor function is given by

$$S(x|\mathcal{F}_t) = \exp\left[-\Lambda(t,t+x)\right], \quad (44)$$

where $S(x|\mathcal{F}_t) := \Pr[(t_{\breve{N}(t)+1} - t) \geq x|\mathcal{F}_t]$. This is the continuous counterpart to the well-known relation between the survivor function and the hazard rate, $S(x_i) = \exp(-\int_0^{x_i} h(u)du)$. Hence, by ignoring the term TK, (43) corresponds to the sum of the conditional survivor function and the conditional intensity function. However, according to Yashin and Arjas (1988), the exponential formula (44) is only valid if $S(x|\mathcal{F}_t)$ is absolutely continuous in $\mathcal{F}_t$, which excludes jumps of the conditional survivor function induced by changes of the information set during a spell. Therefore, in a continuous, dynamic setting, the interpretation of $\exp(-\Lambda(t_{i-1},t_i))$ as a survivor function should be done with caution.

The computation of (43) for a Hawkes model is straightforward. In the case of an exponential decay function, the resulting log likelihood function can be even computed in a recursive way (see e.g. Bowsher (2007)). An important advantage of Hawkes processes is that the individual intensities $\lambda^k(t)$ do not have parameters in common and the parameter vector can be expressed as $\theta = \left(\theta^1, \ldots, \theta^K\right)$, where θ^k denotes the parameters associated with the k-type intensity component. Given that the parameters are variation free, the log likelihood function can be computed as $\ln \mathcal{L}\left(\theta; \{N(t)\}_{t\in(0,T]}\right) = \sum_{k=1}^{K} l^k(\theta^k)$ and can be maximized by maximizing the individual k-type components $l^k(\theta^k)$ separately. This facilitates the estimation particularly when K is large. In contrast, ACI models require to maximize the log likelihood function with respect to all the parameters jointly. This is due to the fact that the ACI innovations are based on the integrated intensities which depend on all individual parameters. The estimation of SCI models is computationally even more demanding since the latent factor has to be integrated out resulting in an n-dimensional integral. Bauwens and Hautsch (2006) suggest to evaluate the likelihood function numerically using the efficient importance sampling procedure introduced by Richard and Zhang (2007). Regularity conditions

for the maximum likelihood estimation of stationary simple point processes are established by Ogata (1981). For more details, see also Bowsher (2007).

Diagnostics for intensity based point process models can be performed by exploiting the stochastic properties of compensators and integrated intensities given in Section 2.4. The model goodness-of-fit can be straightforwardly evaluated through the estimated integrated intensities of the K individual processes, $e^k_{i,1} := \widehat{\Lambda}^k(t^k_{i-1}, t^k_i)$, the integrated intensity of the pooled process $e_{i,2} := \widehat{\Lambda}(t_{i-1}, t_i) = \sum_{k=1}^K \widehat{\Lambda}^k(t_{i-1}, t_i)$, or of the (non-centered) ACI residuals $e_{i,3} := \sum_{k=1}^K \left(\widehat{\Lambda}^k(t^k_{i-1}, t^k_i)\right) y^k_i$. Under correct model specification, all three types of residuals must be i.i.d. standard exponentially distributed. Then, model evaluation is done by testing the dynamic and distributional properties. The dynamic properties are easily evaluated with Portmanteau statistics or tests against independence such as proposed by Brock et al. (1996). The distributional properties can be evaluated using Engle and Russell's (1998) test against excess dispersion (see Section 3.2). Other alternatives are goodness-of-fit tests based on the probability integral transform (PIT) as employed for diagnostics on ACD models by Bauwens et al. (2004).

4.4 Applications

Dynamic intensity models are primarily applied in multivariate financial point processes or whenever a continuous-time setting is particularly required, e.g. to account for time-varying covariates. One strand of applications focusses on the modelling of trading intensities of different types of orders in limit order books. Hall and Hautsch (2007) apply a bivariate ACI model to study the intensities of buy and sell transactions in the electronic limit order book market of the Australian Stock Exchange (ASX). The buy and sell intensities are specified to depend on time-varying covariates capturing the state of the market. On the basis of the buy and sell intensities, denoted by $\lambda^B(t)$ and $\lambda^S(t)$, Hall and Hautsch (2007) propose a measure of the continuous net buy pressure defined by $\Delta^B(t) := \ln \lambda^B(t) - \ln \lambda^S(t)$. Because of the log-linear structure of the ACI model, the marginal change of $\Delta^B(t)$ induced by a change of the covariates is computed as $\gamma^B - \gamma^S$, where γ^B and γ^S denote the coefficients associated with covariates affecting the buy and sell intensity, respectively (see eq. (35)). Hall and Hautsch (2006) study the determinants of order aggressiveness and traders' order submission strategy at the ASX by applying a six-dimensional ACI model to study the arrival rates of aggressive market orders, limit orders as well as cancellations on both sides of the market. In a related paper, Large (2007) studies the resiliency of an electronic limit order book by modelling the processes of orders and cancellations on the London Stock Exchange using a ten-dimensional Hawkes process. Russell (1999) analyzes the dynamic interdependence between the

supply and demand for liquidity by modelling transaction and limit order arrival times at the NYSE using a bivariate ACI model.

Another branch of the literature focusses on the modelling of the instantaneous price change volatility which is estimated on the basis of price durations, see (27) in Section 3.4. This relation is used by Bauwens and Hautsch (2006) to study the interdependence between instantaneous price change volatilities of several blue chip stocks traded at the NYSE based on a SCI model. In this setting, they find a strong evidence for the existence of a common latent component as a major driving force of the instantaneous volatilities on the market. In a different framework, Bowsher (2007) analyzes the two-way interaction of trades and quote changes using a two-dimensional generalized Hawkes process.

References

Aït-Sahalia, Y. (1996): Testing continuous-time models of the spot interest rate. *Review of Financial Studies* **9**, 385–426.

Andersen, T. G. and Bollerslev, T. (1998): Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* **53**, 219–265.

Bauwens, L. and Galli, F. (2008): Efficient importance sampling for ML estimation of SCD models. *Computational Statistics & Data Analysis* to appear.

Bauwens, L., Galli, F., and Giot, P. (2008): The moments of log-ACD models. *Quantitative and Qualitative Analysis in Social Sciences* to appear.

Bauwens, L. and Giot, P. (2000): The logarithmic ACD model: An application to the bid/ask quote process of two NYSE stocks. *Annales d'Economie et de Statistique* **60**, 117–149.

Bauwens, L. and Giot, P. (2001): *Econometric Modelling of Stock Market Intraday Activity*. Kluwer Academic Publishers, Boston, Dordrecht, London.

Bauwens, L. and Giot, P. (2003): Asymmetric ACD models: Introducing price information in ACD models with a two state transition model. *Empirical Economics* **28**, 1–23.

Bauwens, L., Giot, P., Grammig, J., and Veredas, D. (2004): A comparison of financial duration models via density forecasts. *International Journal of Forecasting* **20**, 589–609.

Bauwens, L. and Hautsch, N. (2006): Stochastic conditional intensity processes. *Journal of Financial Econometrics* **4**, 450–493.

Bauwens, L. and Veredas, D. (2004): The stochastic conditional duration model: A latent factor model for the analysis of financial durations. *Journal of Econometrics* **119**, 381–412.

Bisière, C. and Kamionka, T. (2000): Timing of orders, order aggressiveness and the order book at the Paris Bourse. *Annales d'Economie et de Statistique* **60**, 43–72.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T. and Wooldridge, J. (1992): Quasi-maximum likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Reviews* **11**, 143–172.

Bowsher, C. G. (2007): Modelling security markets in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics* **141**, 876–912.

Brémaud, P. and Massoulié, L. (1996): Stability of nonlinear Hawkes processes. *Annals of Probability* **24**, 1563–1588.

Brock, W., Scheinkman, W., Scheinkman, J., and LeBaron, B. (1996): A test for independence based on the correlation dimension. *Econometric Reviews* **15**, 197–235.

Brown, T. C. and Nair, M. G. (1988): A simple proof of the multivariate random time change theorem for point processes. *Journal of Applied Probability* **25**, 210–214.

Cox, D. R. (1972): Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. (1975): Partial likelihood. *Biometrika* **62**, 269.

Daley, D. and Vere-Jones, D. (2003): *An Introduction to the Theory of Point Processes*, volume 1. Springer, New York.

De Luca, G. and Gallo, G. (2004): Mixture processes for financial intradaily durations. *Studies in Nonlinear Dynamics and Econometrics* **8**. Downloadable under http://www.bepress.com/snde/vol8/iss2/art8

De Luca, G. and Zuccolotto, P. (2003): Finite and infinite mixtures for financial durations. *Metron* **61**, 431–455.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998): Evaluating density forecasts, with applications to financial risk management. *International Economic Review* **39**, 863–883.

Drost, F. C. and Werker, B. J. M. (2004): Semiparametric duration models. *Journal of Business and Economic Statistics* **22**, 40–50.

Dufour, A. and Engle, R. F. (2000): The ACD model: Predictability of the time between consecutive trades. *Working Paper, ISMA Centre, University of Reading*.

Durbin, J. and Koopman, S. (2004): Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**, 669–684.

Easley, D. and O'Hara, M. (1992): Time and process of security price adjustment. *The Journal of Finance* **47**, 577–605.

Engle, R. F. (2000): The econometrics of ultra-high-frequency data. *Econometrica* **68, 1**, 1–22.

Engle, R. F. (2002): New frontiers for ARCH models. *Journal of Applied Econometrics* **17**, 425–446.

Engle, R. F. and Lunde, A. (2003): Trades and quotes: A bivariant point process. *Journal of Financial Econometrics* **11**, 159–188.

Engle, R. F. and Ng, V. K. (1993): Measuring and testing the impact of news on volatility. *Journal of Finance* **48**, 1749–1778.

Engle, R. F. and Russell, J. R. (1998): Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–1162.

Engle, R. F. and Russell, J. R. (2005): *Analysis of High Frequency Financial Data. In: Aït-Sahalia, Y. and Hansen, L. (Eds.): Handbook of Financial Econometrics*. North-Holland.

Feng, D., Jiang, G. J., and Song, P. X.-K. (2004): Stochastic conditional duration models with 'leverage effect' for financial tranbsaction data. *Journal of Financial Econometrics* **2**, 390–421.

Fernandes, M. and Grammig, J. (2005): Non-parametric specification tests for conditional duration models. *Journal of Econometrics* **127**, 35–68.

Fernandes, M. and Grammig, J. (2006): A family of autoregressive conditional duration models. *Journal of Econometrics* **130**, 1–23.

Gallant, R. A. (1981): On the bias in flexible functional forms and an essential unbiased form: The Fourier flexible form. *Journal of Econometrics* **15**, 211–245.

Gerhard, F. and Hautsch, N. (2002): Volatility estimation on the basis of price intensities. *Journal of Empirical Finance* **9**, 57–89.

Gerhard, F. and Hautsch, N. (2007): A dynamic semiparametric proportional hazard model. *Studies in Nonlinear Dynamics and Econometrics* **11**. Downloadable under http://www.bepress.com/snde/vol11/iss2/art1

Ghysels, E., Gourieroux, C., and Jasiak, J. (2004): Stochastic volatility duration models. *Journal of Econometrics* **119**, 413–433.

Ghysels, E. and Jasiak, J. (1998): GARCH for irregularly spaced financial data: The ACD-GARCH model. *Studies in Nonlinear Dynamics and Econometrics* **2**, 133–149.
Giot, P. (2005): Market risk models for intraday data. *European Journal of Finance* **11**, 187–212.
Grammig, J. and Wellner, M. (2002): Modeling the interdependence of volatility and inter-transaction duration process. *Journal of Econometrics* **106**, 369–400.
Hafner, C. M. (2005): Durations, volume and the prediction of financial returns in transaction time. *Quantitative Finance* **5**, 145–152.
Hall, A. D. and Hautsch, N. (2006): Order aggressiveness and order book dynamics. *Empirical Economics* **30**, 973–1005.
Hall, A. D. and Hautsch, N. (2007): Modelling the buy and sell intensity in a limit order book market. *Journal of Financial Markets* **10**, 249–286.
Hamilton, J. D. and Jordà, O. (2002): A model of the federal funds rate target. *Journal of Political Economy* **110**, 1135–1167.
Han, A. and Hausman, J. A. (1990): Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**, 1–28.
Hasbrouck, J. (1991): Measuring the information content of stock trades. *Journal of Finance* **46**, 179–207.
Hautsch, N. (2004): *Modelling Irregularly Spaced Financial Data*. Springer, Berlin.
Hautsch, N. (2006): Testing the conditional mean function of autoregressive conditional duration models. *Discussion Paper 2006-06, Finance Research Unit, Department of Economics, University of Copenhagen*.
Hawkes, A. G. (1971): Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58**, 83–90.
Hawkes, A. G. and Oakes, D. (1974): A cluster process representation of a self-exciting process. *Journal of Applied Probability* **11**, 493–503.
Heinen, A. and Rengifo, E. (2007): Multivariate autoregressive modelling of time series count data using copulas. *Journal of Empirical Finance* **14**, 564–583.
Hujer, R. and Vuletic, S. (2007): Econometric analysis of financial trade processes by discrete mixture duration models. *Journal of Economic Dynamics and Control* **31**, 635–667.
Hujer, R., Vuletic, S., and Kokot, S. (2002): The Markov switching ACD model. Finance and Accounting Working Paper, Johann Wofgang Goethe-University, Frankfurt **90**.
Jasiak, J. (1998): Persistence in intratrade durations. *Finance* **19**, 166–195.
Kalbfleisch, J. D. and Prentice, R. L. (1980): *The Statistical Analysis of Failure Time Data*. Wiley, New York.
Karr, A. F. (1991): *Point Processes and their Statistical Inference*. Dekker, New York.
Kiefer, N. M. (1988): Economic duration data and hazard functions. *Journal of Economic Literature* **26**, 646–679.
Koopman, S. J., Lucas, A., and Monteiro, A. (2008): The multi-state latent factor intensity model for credit rating transitions. *Journal of Econometrics* **142**, 399–424.
Koulikov, D. (2002): Modeling sequences of long memory positive weakly stationary random variables. *Technical Report, William Davidson Institute, University of Michigan Business School* **493**.
Lancaster, T. (1997): *The Econometric Analysis of Transition Data*. Cambridge University Press.
Large, J. (2007): Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets* **10**, 1–25.
Liesenfeld, R., Nolte, I., and Pohlmeier, W. (2006): Modelling financial transaction price movements: a dynamic integer count model. *Empirical Economics* **30**, 795–825.
Meddahi, N., Renault, E., and Werker, B. J. (2006): GARCH and irregularly spaced data. *Economics Letters* **90**, 200–204.
Meitz, M. and Saikkonen, P. (2004): Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. *SSE/EFI Working Paper Series in Economics and Finance Stockholm School of Economics* **573**.

Meitz, M. and Teräsvirta, T. (2006): Evaluating models of autoregressive conditional duration. *Journal of Business & Economic Statistics* **24**, 104–124.

Meyer, P. A. (1971): Démonstration simplifiée d'un théorème Knight. In *Lecture Notes in Mathematics* **191**, 191–195. Springer.

Møller, J. and Rasmussen, J. (2004): Perfect simulation of Hawkes processes. *Working Paper, Aalborg University*.

Nelson, D. (1991): Conditional heteroskedasticity in asset returns: A new approach. *Journal of Econometrics* **43**, 227–251.

Ogata, Y. (1981): On Lewis' simulation method for point processes. *IEEE Transactions of Information Theory* **27**, 23–31.

Ogata, Y. (1988): Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* **83**, 9–27.

Prigent, J., Renault, O., and Scaillet, O. (2001): An Autoregressive Conditional Binomial Option Pricing Model. *In: Geman, H., Madan, D., Pliska, S.R. and Vorst, T. (Eds.): Mathematical Finance. Bachelier Congress 2000: Selected Papers from the First World Congress of the Bachelier Finance Society*. Springer Verlag, Heidelberg.

Richard, J. F. and Zhang, W. (2007): Efficient high-dimensional importance sampling. *Journal of Econometrics* **141**, 1385–1411.

Russell, J. R. (1999): Econometric modeling of multivariate irregularly-spaced high-frequency data. *Working Paper, University of Chicago*.

Russell, J. R. and Engle, R. F. (2005): A discrete-state continuous-time model of financial transactions prices and times: The autoregressive conditional multinomial autoregressive conditional duration model. *Journal of Business and Economic Statistics* **23**, 166–180.

Rydberg, T. H. and Shephard, N. (2003): Dynamics of trade-by-trade price movements: Decomposition and models. *Journal of Financial Econometrics* **1**, 2–25.

Strickland, C. M., Forbes, C. S., and Martin, G. M. (2006): Bayesian analysis of the stochastic conditional duration model. *Computational Statistics & Data Analysis* **50**, 2247–2267.

Taylor, S. J. (1986): *Modelling Financial Time Series*. Wiley, New York.

Vere-Jones, D. (1970): Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society, Series B* **32**, 1–62.

Vere-Jones, D. and Ozaki, T. (1982): Some examples of statistical inference applied to earthquake data. *Annals of the Institute of Statistical Mathematics* **34**, 189–207.

Veredas, D., Rodriguez-Poo, J., and Espasa, A. (2002): On the (intradaily) seasonality, dynamics and durations zero of a financial point process. *CORE Discussion Paper 2002/23, Louvain-La-Neuve*.

Yashin, A. and Arjas, E. (1988): A note on random intensities and conditional survival functions. *Journal of Applied Probability* **25**, 630–635.

Zhang, M. Y., Russell, J., and Tsay, R. S. (2001): A nonlinear autoregressive conditional duration model with applications to financial transaction data. *Journal of Econometrics* **104**, 179–207.

Resampling and Subsampling for Financial Time Series

Efstathios Paparoditis and Dimitris N. Politis

Abstract We review different methods of bootstrapping or subsampling financial time series. We first discuss methods that can be applied to generate pseudo-series of log-returns which mimic closely the essential dependence characteristics of the observed series. We then review methods that apply the bootstrap in order to infer properties of statistics based on financial times series. Such methods do not work by generating new pseudo-series of the observed log-returns but by generating pseudo-replicates of the statistic of interest. Finally, we discuss subsampling and self-normalization methods applied to financial data.

1 Introduction

Consider a discrete time process describing the behavior of log-returns

$$R_t = \log\Big(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\Big), \quad t = 1, 2, \ldots$$

where $\{P_t, t = 0, 1, 2, \ldots\}$ is the price of a financial asset observed at time t, t can be measured in seconds, minutes, hours, days, etc. Standard examples for P_t are prices of company-shares quoted at major stock exchanges, interest rates and foreign exchange rates among different currencies. A Taylor series argument shows that R_t is close to the relative returns $(P_t - P_{t-1})/P_{t-1}$ which

Efstathios Paparoditis
Department of Mathematics and Statistics, University of Cyprus, P.O.Box 20537, CY-1678 Nicosia, CYPRUS, e-mail: stathisp@ucy.ac.cy

Dimitris N. Politis
Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA, e-mail: politis@math.ucsd.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,

DOI: 10.1007/978-3-540-71297-8_42,

are free of unit and describe the relative change over time of the price process. Statistical inference for such financial time series has received considerable interest in the last decades leading to a vast and growing literature.

A general class of models proposed to describe the behavior of log returns is given by the following multiplicative model

$$R_t = \mu_t + \sigma_t \cdot \varepsilon_t, \tag{1}$$

where μ_t and σ_t are random variables measurable with respect to the σ-field $\mathcal{F}_{t-1} = \sigma(R_{t-j}, j = 1, 2, \ldots)$ and $\{\varepsilon_t\}$ denotes a sequence of i.i.d. random variables independent of $\{R_{t-j}, j \geq 1\}$ with zero mean and unit variance. It is also assumed that ε_t is independent of the conditional (on the past log-returns $R_{t-j}, j = 1, 2, \ldots$) mean μ_t and of the conditional (non-negative) volatility function σ_t^2. By model (1) the direction of the price change at time t is determined by the sign of ε_t, while the order of magnitude of this change by the volatility process σ_t^2 which is independent of ε_t. In what follows we assume for simplicity that $\mu_t \equiv 0$ and concentrate on bootstrap-based statistical inference for the conditional variance σ_t^2 of the log-returns.

Several of the statistical models proposed in the literature specify the volatility function σ_t^2 as a function of the observable past values R_{t-j} and $\sigma_{t-j}^2, j = 1, 2, \ldots$. For instance, $\{R_t\}$ follows a general nonnegative ARCH(∞) equation if for some known nonnegative function $w(\cdot)$,

$$w(R_t) = \rho_t \xi_t, \tag{2}$$

where

$$\rho_t = a + \sum_{j=1}^{\infty} \beta_j w(R_{t-1}), \tag{3}$$

$\{\xi_t\}$ is a sequence of i.i.d. nonnegative random variables, $a \geq 0$ and $\beta_j \geq 0$, $j = 1, 2, \ldots$; see Robinson (1991), Giraitis et al. (2000) and Kazakevičius and Leipus (2002). $E\xi_1^2 < \infty$ and $\sum_{j=1}^{\infty} \beta_j^2 < 1$ imply weak stationarity of ρ_t. The class (2)-(3) is rich enough and includes as special cases the classical ARCH(p) process (Engle (1982)) as well as the ARCH(∞) process obtained for $w(x) = x^2$,

$$R_t^2 = \Big(a + \sum_{j=1}^{\infty} \beta_j R_{t-j}^2\Big)\varepsilon_t^2.$$

Under certain assumptions on the behavior of the b_j's this class includes also the GARCH(p,q) models, Bollerslev (1986); see also Taylor (1986).

An alternative approach to the above ARCH(∞) model class is the class of stochastic volatility models a simple form of which is given if σ_t in (1) satisfies

$$\sigma_t = g(h_t), \quad \text{with} \quad h_t = a_0 + a_1 h_{t-1} + e_t, \tag{4}$$

where $g(\cdot) > 0$ is a known function and $\{e_t\}$ is an i.i.d. process independent from $\{\varepsilon_t\}$. For $|a_1| < 1$ the process h_t is strictly stationary. Notice that the heteroscedastic variation of R_t described by σ_t^2 is driven by the unobservable latent process $\{h_t\}$ and not by the lagged values of the log-returns $R_{t-1}, R_{t-2}, \ldots$. Thus for this model class, the volatility function is modeled as a strictly stationary process $\{\sigma_t^2\}$ independent of the i.i.d. noise process $\{\varepsilon_t\}$ avoiding therefore any kind of feedback between the noise and the volatility process; cf. Shepard (1996).

Given a time series $R_1, R_2, \ldots, R_n$ of log-returns, one is commonly interested in the construction of point or interval estimators of the volatility function σ_t^2 and in testing hypothesis about this function. Several estimation methods have been proposed in the literature depending on the assumptions imposed on σ_t^2. They rank from parametric methods designed when σ_t^2 belongs to certain parametric classes of models to fully nonparametric methods based on weak assumptions on the function of interest. In this context, nonparametric procedures are useful not only because they lead to estimators of the underlying volatility function without imposing too restrictive assumptions but also because they are very useful for model selection and testing by means of comparing parametric and nonparametric estimates; cf. for instance Kreiss et al. (2008).

Assigning properties of estimators of σ_t^2 is usually carried out by means of asymptotic considerations where the expressions obtained for the asymptotic quantities usually depend in a complicated way on characteristics of the underlying process. This makes alternative approaches based on bootstrap methodology appealing. During the last decades different bootstrap methods have been proposed in the context of financial time series. Some early application of the bootstrap in financial time series mainly based on i.i.d. resampling (cf. Maddala and Li (1996) and Ruiz and Pascual (2002) for a review) are not appropriate and may lead to wrong conclusions since log-returns are not independent despite their vanishing correlation. Thus approaches to bootstrap financial time series should take into account their dependence structure or at least those aspects of this dependence structure which are important for the particular inference problem at hand.

Nonparametric methods to bootstrap time series which are based on resampling with replacement from blocks of consecutive observations can in principle be applied to financial series, and their properties to approximate the distribution of statistics of interest can be investigated provided the underlying stochastic process obeys some appropriate weak dependence, e.g., mixing conditions. For mixing properties of some commonly used time series models with applications in finance, see for instance, Carrasco and Chen (2002). Under appropriate mixing conditions, bootstrapping a series of log-returns $R_1, R_2, \ldots, R_n$ can be done by randomly choosing with replacement a number of l, $l = [n/b]$, blocks of b consecutive values $\{R_t, R_{t+1}, \ldots, R_{t+b-1}\}$ from all possible $n - b$ blocks; cf. Künsch (1989), Liu and Sign (1992), Politis and Romano (1994); see also Bühlmann (2002) and Härdle et al. (2003) for an

overview and the monograph by Lahiri (2003). Although such general blocking techniques preserve the dependence structure of the observations within blocks, they have not been widely used in the context of financial time series. A reason for this might be that since it is common to impose some kind of model structure in describing the behavior of financial time series, efficiency considerations make model-based bootstrap methods more attractive. Furthermore, and concerning nonparametric estimators of the volatility function σ_t^2, it is well-known that the dependence structure of the underlying process affects the limiting behavior of the estimators only through the behavior of certain finite dimensional, stationary distributions of the process. This again suggests that bootstrapping nonparametric estimators for financial time series can be successfully done without mimicking the whole and probably very complicated dependence structure of the observed series.

In Section 2 we review different methods to bootstrap financial time series, that is methods that can be applied to generate pseudo-replicates $R_1^*, R_2^*, \ldots, R_n^*$ of the observed series of log-returns. Such methods are designed in a way that mimics closely the essential dependence characteristics of the observed time series or at least those characteristics which are important for inferring consistently properties of the statistics of interest. In Section 3 we concentrate on the somewhat different problem on how to apply the bootstrap in order to infer properties of statistics based on financial time series. Such methods do not work by generating new pseudo-observations of the observed log-returns that preserves their dependence structure, but by generating pseudo-replicates of the statistic of interest. Section 4 is devoted to subsampling and self-normalization methods applied to financial data.

2 Resampling the Time Series of Log–Returns

The bootstrap procedures described in this section generate replications $R_1^*, R_2^*, \ldots, R_n^*$ of the series of log-returns. Bootstrap-based inference is then provided by approximating properties of the statistics based on the original time series $R_1, R_2, \ldots, R_n$ by the corresponding properties of the statistics based on the bootstrap replicates $R_1^*, R_2^*, \ldots, R_n^*$.

2.1 Parametric methods based on i.i.d. resampling of residuals

The basic idea to bootstrap a series of log-returns when parametric assumptions on the volatility function $\sigma_t^2(\cdot)$ are imposed, is to obtain residuals using the estimated parametric model and to generate new pseudo-series of log-returns using the estimated model structure and i.i.d. resampling of residuals.

More specifically, suppose that

$$R_t = \sigma_t(\theta)\varepsilon_t \tag{5}$$

where the function $\sigma_t^2(\theta)$ belongs to some parametric family of functions with $\theta \in \Theta$ and Θ a finite dimensional parameter space. Furthermore, $\{\varepsilon_t\}$ denotes a sequence of i.i.d. random variables with zero mean and unit variance.

As an example consider a GARCH(p,q) specification of $\sigma_t^2(\theta)$ given by

$$\sigma_t^2(\theta) = c + \sum_{i=1}^{p} a_i R_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2, \tag{6}$$

$\theta = (c, a_1, \ldots, a_p, b_1, \ldots, b_q)^{'}$, $c > 0$, $a_i \geq 0$, $b_j \geq 0$ and p and q nonnegative integers. Strictly stationarity of the stochastic process $\{X_t, t \in Z\}$ with $E(X_t^2) < \infty$ follows if $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$, Bollerslev (1986); see also Bougerol and Picard (1992). Let $\{\widehat{\theta} = \widehat{\theta}_n, n \in N\}$ be a sequence of estimators $\widehat{\theta} = (\widehat{c}, \widehat{a}_1, \ldots, \widehat{a}_p, \widehat{b}_1, \ldots, \widehat{b}_q)^{'}$ of θ, for instance, the commonly used conditional maximum likelihood estimator. Define standardized residuals

$$\widehat{\varepsilon}_t = \frac{\widetilde{\varepsilon}_t - \widetilde{n}^{-1} \sum_i \widetilde{\varepsilon}_i}{\{\widetilde{n}^{-1} \sum_i \widetilde{\varepsilon}_i^2 - (\widetilde{n}^{-1} \sum_i \widetilde{\varepsilon}_i)^2\}^{1/2}},$$

where $\widetilde{\varepsilon}_t = R_t / \sigma_t(\widehat{\theta})$ are the estimated model residuals and $\widetilde{n} = n - p$. Pseudo-series of log returns following a GARCH(p,q) model structure can then be generated using the equation

$$R_t^* = \sigma_t(\widehat{\theta})\varepsilon_t^*,$$

where $\sigma_t(\widehat{\theta})$ is the specification (6) of the conditional variance function with θ replaced by its estimator $\widehat{\theta}$, $\{\varepsilon_t^*\}$ is an i.i.d. sequence with $\varepsilon_t^* \sim F_n$ and F_n is the empirical distribution function of the $\widehat{\varepsilon}_t$'s.

In most applications of such a parametric bootstrap procedure it is assumed that ε_t has finite fourth moments, i.e., $E\varepsilon_t^4 < \infty$. For instance and for ε_t being standard Gaussian errors, Kokoszka et al. (2004) use this bootstrap method to construct confidence intervals for the autocorrelations of the squares of log-returns, while Miguel and Olave (1999) considered parametric bootstrap prediction intervals for ARCH processes. The assumption of finite fourth moments is crucial for proving consistency of such a parametric bootstrap, since, loosely speaking, this assumption ensures asymptotic normality of the estimator involved with an appropriate, $\sqrt{n}$-rate of convergence. Without such a moment assumption consistency of this parametric bootstrap procedure might be questionable.

For instance, let $\widehat{\theta}^*$ be the estimator of θ based on the pseudo-returns $R_1^*, R_2^*, \ldots, R_n^*$ and consider the problem of estimating the distribution of an appropriately rescaled version of $\widehat{\theta} - \theta$ by the corresponding distribution of the

bootstrap estimator $\widehat{\theta}^* - \widehat{\theta}$. In such a context, consistency of the parametric bootstrap procedure depends on the limiting behavior of $\widehat{\theta} - \theta$ which in turn depends on the distribution of the i.i.d. errors ε_t. If $E(\varepsilon_1)^4 < \infty$ then it has been shown by Berkes et al. (2003) that under some regularity conditions, $\sqrt{n}(\widehat{\theta}-\theta) \Rightarrow N(0, (E\varepsilon_1^4-1)\Sigma_\theta^{-1})$ as $n \to \infty$, where $\Sigma_\theta = E(\sigma_1^{-4}(\theta)U(\theta)U^{'}(\theta))$ and $U(\theta)$ is the r-dimensional vector of first derivatives of $\sigma_1^2(\theta) = \sigma_1^2(a, b, c)$ with respect to the components of $a = (a_1, \ldots, a_p), b = (b_1, \ldots, b_q)$ and c, evaluated at θ. The asymptotic normality of $\sqrt{n}(\widehat{\theta} - \theta)$ in this case, suggests the use of the distribution of $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ to approximate the distribution of the former statistic. However, if $E(\varepsilon_1^4) = \infty$ then the normalizing sequence of the estimation error $\widehat{\theta} - \theta$ is no more $\sqrt{n}$ and the limiting distribution of the appropriately normalized sequence $\widehat{\theta} - \theta$ differs from a Gaussian law and depends on the distribution of the squared errors ε_1^2; see Straumann (2005) and Hall and Yao (2003). In such situations, subsampling offers an alternative; cf. Politis et al. (1999). A subsampling based approach to infer properties of parameter estimators for GARCH processes which is asymptotically valid under general assumptions on the distribution of the errors ε_t has been proposed by Hall and Yao (2003); see Section 4 for details.

2.2 Nonparametric methods based on i.i.d. resampling of residuals

If instead of a parametric structure for $\sigma(\cdot)$ the general nonparametric model

$$R_t = \sigma(R_{t-1}, R_{t-2}, \ldots, R_{t-p})\varepsilon_t, \tag{7}$$

is assumed, then the volatility function $\sigma^2(\cdot)$ can be estimated in a nonparametric way and pseudo-series of log-returns can be generated using the estimated volatility function and i.i.d. resampling of estimated residuals. Meaningful statistical inference for $\sigma^2(\cdot)$ in a nonparametric context, requires however, that the discrete time Markov process $\{R_t\}$ following (7) satisfies some stability and ergodicity conditions which can be achieved by imposing some restrictions on the distribution of the errors ε_t and on the function $\sigma : R^p \to (0, \infty)$. For instance, when $p = 1$, $\{R_t\}$ is geometrically ergodic if the density f_ε of ε is absolutely continuous and positive everywhere, $\inf_{x\in R} \sigma(x) > 0$, σ is bounded over bounded sets and satisfies $\limsup_{|x|\to\infty} E|\sigma(x)\varepsilon_1|/|x| < 1$; cf. Franke et al. (2002b).

Let $\widehat{\sigma}^2(x_1, \ldots, x_p)$ be a nonparametric estimator of $\sigma^2(x_1, \ldots, x_p) = E(R_t^2|R_{t-1} = x_1, \ldots, R_{t-p} = x_p)$ and define standardized residuals

$$\widehat{\varepsilon}_t = \frac{\widetilde{\varepsilon}_t - \widetilde{n}^{-1}\sum_i \widetilde{\varepsilon}_i}{\{\widetilde{n}^{-1}\sum_i \widetilde{\varepsilon}_i^2 - (\widetilde{n}^{-1}\sum_i \widetilde{\varepsilon}_i)^2\}^{1/2}},$$

where $\widetilde{\varepsilon}_t = R_t/\widehat{\sigma}(R_{t-1}, R_{t-2}, \ldots, R_{t-p})$, $t = p+1, p+2, \ldots, n$ and $\widetilde{n} = n-p$. Pseudo-series of log returns can then be generated as

$$R_t^* = \widehat{\sigma}(R_{t-1}^*, \ldots, R_{t-p}^*)\varepsilon_t^*, \tag{8}$$

where the ε_t^*'s are i.i.d. random variables with $\varepsilon_t^* \sim F_n$ and F_n is the empirical distribution function of the $\widehat{\varepsilon}_t$; cf. Franke et al. (2002a).

To fix ideas, let $p = 1$ and consider the following nonparametric estimator of the volatility function

$$\widehat{\sigma}_h^2(x) = \frac{1}{\widehat{f}_{R_t}(x)} \frac{1}{(n-1)} \sum_{t=1}^{n-1} K_h(x - R_t)(R_{t+1} - \overline{R}_n)^2, \tag{9}$$

where h is a smoothing bandwidth, K a smoothing kernel and $\overline{R}_n = n^{-1}\sum_{t=1}^{n} R_t$. Notice that centering by the sample mean $\overline{R}_n$ is used since it is assumed that $\mu_t = 0$. If the conditional mean function μ_t is not constant and $\mu_t = m(R_{t-1})$ with $m(\cdot)$ some smooth function, then $\overline{R}_n$ should be replaced by a nonparametric estimator $\widehat{m}_h(R_{t-1})$ of the conditional mean function $m(R_{t-1}) = E(R_t|R_{t-1})$; see Fan and Yao (1998) for details and for a comparison of different estimators of the volatility function.

The bootstrap procedure described above can be applied to approximate the distribution of random variables like $\sqrt{nh}(\widehat{\sigma}_h^2(x) - \sigma^2(x))$ and $\sup_{x\in[a,b]} |\widehat{\sigma}_h^2(x) - \sigma^2(x)|$, for some $a < b$, or of some standardized versions thereof. Franke et al (2002b) established under certain regularity conditions absolute regularity and geometric ergodicity of the bootstrap process $\{R_t^*, t \in Z\}$ and applied these results to show validity of the corresponding nonparametric bootstrap procedure in approximating the distribution of supremum type statistics of the conditional mean estimator. Franke et al. (2000a) consider applications to pointwise statistics while Franke et al. (2004) investigated properties of such a bootstrap procedure for the construction of confidence bands for the volatility function.

Notice that an inherent problem common in applications of the bootstrap to estimate the distribution of nonparametric estimators is how to deal with the bias. In particular, and decomposing the statistic of interest in a stochastic and in a bias term, i.e., writing

$$\sqrt{nh}(\widehat{\sigma}_h^2(x) - \sigma^2(x)) = \sqrt{nh}(\widehat{\sigma}_h^2(x) - E(\widehat{\sigma}_h^2(x)) + \sqrt{nh}(E(\widehat{\sigma}_h^2(x)) - \sigma^2(x)),$$

makes it clear that if one is interested in applying the bootstrap in order to approximate the distribution of $\sqrt{nh}(\widehat{\sigma}_h^2(x) - \sigma^2(x))$, then such a bootstrap procedure should approximate correctly the behavior of both terms on the right hand side of the above decomposition. In this context correct estimation of the bias term requires a kind of oversmoothing, i.e., the use of an estimator $\widehat{\sigma}_g^2(x)$ to generate the pseudo-variables R_t^*'s in (8) for which the bandwidth g satisfies $g > h$ and $h/g \to 0$ as $n \to \infty$. The bootstrap

statistic used to approximate the distribution of $\sqrt{nh}(\widehat{\sigma}_h^2(x) - \sigma^2(x))$ is then given by $\sqrt{nh}(\widehat{\sigma}_h^{*^2}(x) - \sigma_g^2(x))$ where $\widehat{\sigma}_h^{*^2}(x)$ is the estimator (9) with R_t replaced by the bootstrap pseudo-variables R_t^* and $\overline{R}_n$ by $\overline{R}_n^* = n^{-1}\sum_{t=1}^n R_t^*$. An alternative to such a bootstrap estimation of the bias would be explicit bias correction; see Hall(1992) for a discussion of this issue in the context of independent data.

Although the above nonparametric bootstrap procedure can be successfully applied to estimate the distribution of the nonparametric estimators of interest, it fails if the process of log-returns does not follow model (7). The reason for this is that in this case the stationary distribution of the bootstrap variables $(R_t^*, R_{t-1}^*, \ldots, R_{t-p}^*)$ does not converge to the desired stationary distribution of $(R_t, R_{t-1}, \ldots, R_{t-p})$. In fact and for $\mathbf{R}_{t-1,t-p} = (R_{t-1}, R_{t-2}, \ldots, R_{t-p})$, we expect is such a case that the conditional distribution function $P(R_t^* \leq \cdot | \mathbf{R}_{t-1,t-p}^* = \mathbf{x})$ of the bootstrap process will behave asymptotically like $F_{U_t}(\cdot/\sigma(\mathbf{x}))$ where F_{U_t} denotes the stationary distribution function of $U_t = R_t/\sigma(\mathbf{x})$ and $\sigma(\mathbf{x}) = \sqrt{\sigma^2(\mathbf{x})}$ with $\sigma^2(\mathbf{x}) = E(R_t^2 | R_{t-i} = x_i, i = 1, 2, \ldots, p)$. This distribution differs however from the conditional distribution function $P(R_t \leq \cdot | \mathbf{R}_{t-1,t-p} = \mathbf{x})$ of the underlying process if model (7) is not correct.

2.3 Markovian bootstrap

Model (7) is a special case of a more general p-th order Markovian process, that is a process $\{R_t, t \in Z\}$ which satisfies

$$P(R_t \in A | \sigma(R_s, s < t)) = P(R_t \in A | R_{t-1}, R_{t-2}, \ldots, R_{t-p}),$$

for all $A \in \mathcal{B}(R)$ and all $t \in Z$. Suppose that such a model describes the behavior of log-returns and denote by $F_{R_t|\mathbf{R}_{t-1,t-p}}(\cdot \,|\, \mathbf{x}) = P(R_t \leq y | R_{t-i} = x_i, i = 1, 2, \ldots, p)$ the one-step transition distribution function. Assume that the corresponding conditional probability measure possesses a density with respect to Lebesgue measure which we denote by $f_{R_t|\mathbf{R}_{t-1,t-p}}(\cdot|\mathbf{x})$. Notice that model (7) is a specific Markov process with one-step transition distribution function given by $F_{R_t|\mathbf{R}_{t-1,t-p}}(y|\mathbf{x}) = F_\varepsilon(y/\sigma(\mathbf{x}))$ where F_ε is the distribution function of the error ε_1.

Imposing conditions on $F_{R_t|\mathbf{R}_{t-1,t-p}}(\cdot)$ which ensure stationarity and geometric ergodicity of the associated Markov chain, bootstrap replicates of log-returns can be generated using the Markovian model structure without specifying its functional form. Rajarshi (1990) proposed such a bootstrap approach based on a nonparametric estimator $\widehat{f}_{R_t|\mathbf{R}_{t-1,t-p}}(\cdot|\mathbf{x})$ of the one step transition density $f_{R_t|\mathbf{R}_{t-1,t-p}}(\cdot|\mathbf{x})$. Such a nonparametric estimator is for instance given by

$$\widehat{f}_{R_t|\mathbf{R}_{t-1,t-p}}(y|\mathbf{x}) = \sum_{t=p+1}^{n} K_b((y,\mathbf{x}) - \mathbf{R}_{t,t-p}) \Big/ \sum_{t=p+1}^{n} K_b(\mathbf{x} - \mathbf{R}_{t-1,t-p}),$$

where b is the bandwidth used to estimate the stationary densities of interest. New series of pseudo-replications can then be generated as

$$R_t^* \sim \widehat{f}_{R_t|\mathbf{R}_{t-1,t-p}}(\cdot|R_{t-1}^*, R_{t-2}^*, \ldots, R_{t-p}^*).$$

A different approach which does not require explicit nonparametric estimation and resamples directly the original series of log-returns in an appropriate way preserving their Markovian dependence structure, has been proposed by Paparoditis and Politis (2001a). Given a series $R_{t-p}^*, R_{t-p+1}^*, \ldots, R_{t-1}^*$ of pseudo log-returns, their approach works by generating a new pseudo-variable R_t^* as

$$R_t^* = R_J,$$

where J is a discrete random variable taking values in the set $\{p+1, p+2, \ldots, n\}$ with

$$P(J = s) = \frac{W_b(\mathbf{R}_{t-1,t-p}^* - \mathbf{R}_{s-1,s-p})}{\sum_{l=p+1}^{n} W_b(\mathbf{R}_{t-1,t-p}^* - \mathbf{R}_{l-1,l-p})}$$

for $s \in \{p+1, p+2, \ldots, n\}$, where $\mathbf{R}_{t-1,t-p}^* = (R_{t-p}^*, R_{t-p+1}^*, \ldots, R_{t-1}^*)$. Here $W_b(\cdot) = b^{-p}W(\cdot/p)$, where b is the so-called resampling width and $W(\cdot)$ a p-dimensional, nonnegative and symmetric resampling kernel with mean zero. Notice that this procedure resamples the observed log-returns in a way according to which the probability of R_s being selected is higher the closer is its preceding segment $(R_{s-1}, R_{s-2}, \ldots, R_{s-p})$ to the last generated bootstrap segment $(R_{t-1}^*, R_{t-2}^*, \ldots, R_{t-p}^*)$.

Properties of Markovian bootstrap procedures have been investigated by Rajarshi (1990) and Paparoditis and Politis (2001a); see also Horowitz (2003). Applications of such a bootstrap procedure in order to approximate the distribution of nonparametric conditional moment estimators and for constructing pointwise confidence intervals have been investigated by Paparoditis and Politis (2002). Notice that in approximating correctly the bias of the nonparametric estimators involved, a kind of oversmoothing condition is needed here as well leading to some restrictions on the behavior of the resampling width b compared to the smoothing bandwidth h. In particular, b should satisfy $b > h$ and $b/h \to 0$ as $n \to \infty$.

We stress here the fact that for the kind of nonparametric estimators discussed in this paper, the range of applicability of the above Markovian procedures goes far beyond the Markov process class. This is due to the fact that the Markovian resampling schemes described above mimics correctly the $(p+1)$-dimensional stationary distribution of $(R_t, R_{t-1}, \ldots, R_{t-p})$ even if the underlying process is not Markov. This property suffices to establish consistency of the above Markovian bootstrap procedures applying to estimate the

distribution of the nonparametric estimators of interest for a very broad class of stochastic processes; see Paparoditis and Politis (2002) for details.

3 Resampling Statistics Based on the Time Series of Log–Returns

In the context of dependent data it is possible in certain situations to apply the bootstrap to some statistics of interest without generating new time series of pseudo-observations that preserve the dependence structure of the observed time series. Such applications resample directly the statistic of interest in a way which mimics correctly those characteristics of the dependence structure of the underlying process which are essential for the random behavior of the statistic of interest.

Suppose for instance, that we are interested in approximating the poinwise distribution of $\sqrt{nh}(\widehat{\sigma}_h^2(x)-\sigma^2(x))$ by means of the bootstrap. As it has been already stressed, for a bootstrap procedure to be successful in approximating correctly the (limiting) distribution of this statistic, it is not necessary to imitate the whole and probably very complicated dependence structure of the underlying process of log-returns. For this, it suffices to mimic correctly the $(p+1)$-dimensional stationary distribution of $(R_t, R_{t-1}, \ldots, R_{t-p})$. This is a consequence of the basic fact that the asymptotic distribution of nonparametric estimators does not reflect the dependence structure of the underlying process beyond the $(p+1)$-dimensional structure; cf. Robinson (1983). Hart (1995) called this the whitening by windowing effect. This basic observation has motivated the development of bootstrap procedures that generate pseudo-replicates $(R_t^*, R_{t-1}^*, \ldots, R_{t-p}^*)$ of $(R_t, R_{t-1}, \ldots, R_{t-p})$ in a way that mimic correctly the joint distribution of the last random vector, without generating new pseudo-series of log-returns.

3.1 Regression bootstrap

The regression bootstrap is a nonparametric bootstrap procedure which generates replicates of the pairs $\{(R_t, \mathbf{R}_{t-1,t-p}), t = p+1, p+2, \ldots, n\}$ denoted by $\{(R_t^*, \mathbf{R}_{t-1,t-p}), t = p+1, p+2, \ldots, n\}$ by using a fixed design, heteroscedastic regression model with errors obtained by i.i.d. resampling from estimated model residuals; cf. Franke et al. (2002a). Notice that in this resampling scheme only the random variable R_t is bootstrapped while $\mathbf{R}_{t-1,t-p}$ is treated as a (conditionally) fixed design. In particular, the bootstrap variables R_t^* are generated using the equation

$$R_t^* = \widehat{\sigma}_g(R_{t-1}, R_{t-2}, \ldots, R_{t-p})\varepsilon_t^*, \tag{10}$$

where the ε_t^* are independent random variables such that $\varepsilon_t^* \sim F_n$ and F_n is the empirical distribution function of the estimated errors $\widehat{\varepsilon}_t$ given in Section 2.2. Notice that the random variables R_t^* are (conditional on the observed series) independent with $E^*(R_t^*) = 0$ and $Var^*(R_t^*) = \widehat{\sigma}_g^2(R_{t-1}, R_{t-2}, \ldots, R_{t-p})$. Thus the dependence structure of the series of log returns is not preserved by this bootstrap method. The distribution of $\sqrt{nh}(\widehat{\sigma}_h^2(\mathbf{x}) - \sigma^2(\mathbf{x}))$ can now be approximated by that of $\sqrt{nh}(\widehat{\sigma}_h^{*^2}(\mathbf{x}) - \widehat{\sigma}_g^2(\mathbf{x}))$, where, for $p = 1$,

$$\widehat{\sigma}_h^{*^2}(x) = \frac{1}{\widehat{f}_{R_t}(x)} \frac{1}{(n-1)} \sum_{t=1}^{n-1} K_h(x - R_t)(R_{t+1}^* - \overline{R}_n^*)^2,$$

and $\overline{R}_n^* = n^{-1} \sum_{t=1}^n R_t^*$. The bandwidth g used in (10) can be chosen so that the above bootstrap procedure estimates also correctly the bias term $\sqrt{nh}(E(\widehat{\sigma}_h^2(x)) - \sigma^2(x))$ of the nonparametric volatility function estimator. For this an oversmoothing type condition should be satisfied, i.e., $g > h$ with $h/g \to 0$ as $n \to \infty$; cf. Franke et al. (2002a).

3.2 Wild bootstrap

The wild bootstrap methodology can be also applied in the context of financial time series to estimate the distribution of some nonparametric estimators of interest; see Franke et al. (2002a) and Kreiss (2000). To fix ideas, let $p = 1$ and consider the centered log-returns $Y_t = R_t - \overline{R}_n$, $t = 1, 2, \ldots, n$. Let further η_t, $t = 1, 2, \ldots, n$ be a sequence of independent, identically distributed random variables satisfying $E(\eta_t) = 0$, $E(\eta_t^2) = 1$. For higher order performance the distribution of η_t is often chosen such that additionally the condition $E(\eta_t^3) = 1$ is satisfied; cf. Mammen (1992) for a discussion.

The wild bootstrap works by generating pairs $\{(Y_{t+1}^*, R_t), t = 1, 2, \ldots, n-1\}$ where

$$Y_{t+1}^* = \widehat{\sigma}_g^2(R_t) + \varepsilon_{t+1}^*, \tag{11}$$

and

$$\varepsilon_{t+1}^* = \left[Y_{t+1}^2 - \widehat{\sigma}_h^2(R_t)\right] \cdot \eta_{t+1} \tag{12}$$

Notice that $E^*(Y_{t+1}^*) = \widehat{\sigma}_g^2(R_t)$ and $Var^*(Y_{t+1}^*)^2 = (Y_{t+1}^2 - \widehat{\sigma}_h^2(R_t))^2$. Furthermore, the bootstrap random variables Y_{t+1}^* are (conditional on the observed sample) independent, i.e., the dependence structure of the observed time series is not preserved by this bootstrap scheme. In fact, in the bootstrap world the Y_{t+1}^*'s are generated according to a fixed design nonparametric regression model with mean $\widehat{\sigma}_g^2(R_t)$ and variance $Var^*(\varepsilon_{t+1}^*)$. Now, to approximate the distribution of $\sqrt{nh}(\widehat{\sigma}_h^2(x) - \sigma^2(x))$ the wild-bootstrap statistic $\sqrt{nh}(\widehat{\sigma}_h^{*^2}(x) - \sigma_g^2(x))$ can be used, where

$$\widehat{\sigma}_h^{*^2}(x) = \frac{1}{\widehat{f}_{R_t}(x)} \frac{1}{(n-1)} \sum_{t=1}^{n-1} K_h(x - R_t)(Y_{t+1}^*)^2.$$

Notice that in order to capture correctly also the bias term of the above nonparametric estimation, and similar to all bootstrap approaches discussed so far, an oversmoothing type condition is needed by the wild bootstrap as well; see Franke et al. (2002a) and Kreiss (2000).

Although the wild bootstrap does not preserve the dependence structure of the observed series, it resamples correctly the distribution of the nonparametric statistics of interest. This bootstrap scheme is robust against model misspecifications, at least as far as the estimation of the distribution of pointwise statistics like $\sqrt{nh}(\widehat{\sigma}_h^2(\mathbf{x}) - \sigma^2(\mathbf{x}))$ is concerned; cf. Kreiss (2000). Neumann and Kreiss (1998) applied a wild bootstrap to the construction of uniform confidence bands for the conditional mean based on supremum type statistics and using strong approximation results under the assumption of a Markovian model. Kreiss (2000) considered the problem of estimating by means of a wild bootstrap procedure the distribution of the sup-distance involved in the construction of simultaneous confidence intervals for the volatility function. His approach is also based on the assumption that the underlying process obeys a Markovian dependence structure.

3.3 Local bootstrap

Another simple and totally model free way to bootstrap nonparametric estimators in time series is the so-called local bootstrap; see Shi (1991) for the case of i.i.d. data and Paparoditis and Politis (2000) for the case of dependent data. This bootstrap method generates replicates $\{(R_t^*, \mathbf{R}_{t-1,t-p}); t = p+1, p+2, \ldots, n\}$ of the observed pairs $\{(R_t, \mathbf{R}_{t-1,t-p}); t = p+1, p+2, \ldots, n\}$ by correctly imitating the conditional distribution $F_{R_t,\mathbf{R}_{t-1,t-p}}(\cdot|\mathbf{x})$ of the log-returns. In contrast to the regression bootstrap, the local bootstrap resamples the observed values R_t by giving more resampling weights to the values R_s for which $\mathbf{R}_{s-1,s-p}$ is close to $\mathbf{R}_{t-1,t-p}$.

For $p = 1$ and for the problem of estimating the distribution of the nonparametric estimator (9) of the conditional variance function, this procedure can be described as follows. Let $p = 1$ and $Y_t = R_t - \overline{R}_n$, $t = 1, 2, \ldots, n$. Bootstrap replicates $\{(Y_{t+1}^*, R_t), t = 1, 2, \ldots, n-1\}$ of the pairs $\{(Y_{t+1}, R_t), t = 1, 2, \ldots, n-1\}$ are then generated so that the bootstrap pseudo-variable Y_{t+1}^* satisfies

$$P(Y_{t+1}^* = Y_{s+1}|R_t) = \frac{W_b(R_t - R_s)}{\sum_{l=1}^{n-1} W_b(R_t - R_l)}.$$

Here b is the so-called resampling width which determines the neighborhood from which replicates of Y_{t+1} are selected and $W(\cdot)$ a resampling kernel. It

is easily seen that

$$E^*((Y^*_{t+1})^2|R_t = x) = \frac{1}{\widehat{f}_{R_t,b}(x)} \frac{1}{(n-1)} \sum_{t=1}^{n-1} W_b(x - R_t)(R_{t+1} - \overline{R}_n))^2,$$

i.e., $E^*((Y^*_{t+1})^2|R_t = x) = \widehat{\sigma}^2_b(x)$. This suggests that the distribution of $\sqrt{nh}(\widehat{\sigma}^2_h(x) - \sigma^2(x))$ can be approximated by that of $\sqrt{nh}(\widehat{\sigma}^{*^2}_h(x) - \widehat{\sigma}^2_b(x))$ where

$$\widehat{\sigma}^{*^2}_h(x) = \frac{1}{\widehat{f}_{R_t,h}(x)} \frac{1}{(n-1)} \sum_{t=1}^{n-1} K_h(x - R_t)(Y^*_{t+1})^2.$$

Consistency properties of such a bootstrap procedure for estimating the distribution of pointwise statistics based on nonparametric estimators of conditional moments for time series data have been established by Paparoditis and Politis (2000). Ango Nze et al. (2002) used such a procedure to estimate the distribution of nonparametric moment estimators under weak dependent assumptions (see Doukhan and Louchichi (1999)).

4 Subsampling and Self–Normalization

Subsampling for dependent data is a method valid in extreme generality. To define it, let $\widehat{\theta}_n = \widehat{\theta}_n(R_1, \ldots, R_n)$ be an arbitrary statistic that is consistent for a general parameter θ at rate a_n, i.e., for large n, the law of $a_n(\widehat{\theta}_n - \theta)$ tends to some well-defined asymptotic distribution J. The rate a_n does not have to equal $\sqrt{n}$, and the distribution J does not have to be normal; we do not even need to know its shape, just that it exists. Let $\widehat{\theta}_{i,b} = \widehat{\theta}_b(R_i, \ldots, R_{i+b-1})$ be the subsample value of the statistic computed from the ith block of length b. The subsampling estimator of J is $\widehat{J}_{b,SUB}$ defined as the empirical distribution of the normalized (and centered) subsample values $a_b(\widehat{\theta}_{i,b} - \widehat{\theta}_n)$ for $i = 1, \ldots, q$ where $q = n - b + 1$.

The asymptotic consistency of $\widehat{J}_{b,SUB}$ for general statistics under very weak conditions was shown in Politis and Romano (1994); consequently, confidence intervals and/or tests for θ can immediately be formed using the quantiles of $\widehat{J}_{b,SUB}$ instead of the quantiles of the (unknown) J. Notice that if only a variance estimator for $a_n\widehat{\theta}_n$ is sought, it can be constructed by the sample variance of the normalized subsample values $a_b\widehat{\theta}_{i,b}$ for $i = 1, \ldots, q$; consistency of the subsampling estimator of variance was shown by Carlstein (1986) under some uniform integrability conditions.

In addition to the usual requirement $b \to \infty$ as $n \to \infty$ but with $b/n \to 0$, the conditions of Politis and Romano (1994) boil down to:

C1 The series $\{R_t\}$ is strictly stationary.
C2 The series $\{R_t\}$ is strong mixing.

C3 The rate a_n is known.

Notice that condition C3 is important for the practical construction of confidence intervals and tests, not for the consistency of subsampling as a method.

Although Conditions C1-C3 are quite weak, they manage to exclude a number of interesting settings pertaining to financial time series since the latter are often plagued by heteroscedasticity, long-range dependence, and heavy tails. Fortunately, all the above conditions can still be relaxed. A review of non-standard conditions for which (variations of) the block bootstrap are consistent can be found in Politis (2003) together with a discussion on the important issue of block size choice; see also the monograph by Lahiri (2003). Condition C1 can easily be relaxed to just asymptotic stationarity as enjoyed, for example, by Markov processes that are generated with an arbitrary (non-equilibrium) start-up distribution; see Politis et al. (1999, Ch. 4). The strong mixing condition C2 has been recently relaxed to the weak dependence condition of Doukhan and Louhichi (1999); see Ango Nze et al. (2003). Finally, condition C3 was relaxed since Bertail et al. (1999) showed how to construct subsampling-based estimators of the rate a_n that can be employed for the construction of confidence intervals and tests.

The above can be seen as small improvements/perturbations on the original Conditions C1-C3. We now describe in more detail how to address major break-downs of those conditions using two broad techniques: (a) examination of the time series of subsample values $\widehat{\theta}_{i,b}$ for $i = 1, \ldots, q$, and (b) the idea of self-normalization.

To elaborate on using technique (a) in order to relax Condition C2, consider the familiar situation of a unit-root test such as the set-up in Paparoditis and Politis (2003). Under the unit-root hypothesis the data are not strong mixing, not even weakly dependent; however, the subsample values of the Phillips-Perron (say) unit-root statistic *are* weakly dependent and thereby the validity of subsampling is preserved; see Politis et al. (1999, Theorem 12.2.1), Politis et al. (2004), and Romano and Wolf (2001).

The same idea can be applied to relax Condition C1. For example, consider a series that is not strictly stationary, e.g., heteroskedastic. If the subsample values of the pertinent statistic have distributions that converge to the limit J in a uniform way, then subsampling remains consistent; see Politis et al. (1997). An important example is least squares regression with dependent errors and/or covariates; subsampling is shown to work here under conditions *without* assuming a stationary or homoskedastic error structure (Politis et al. 1997, Theorem 3.4).

The idea of self-normalization is closely related to that of studentization. To describe it, we now focus on a particular example that may defy Condition C3: assume Condition C1, and consider the simple case where the statistic of interest $\widehat{\theta}_n = n^{-1} \sum_{t=1}^{n} R_t$ is the sample mean which is an estimator of the expected value $\theta = E(R_t)$ (assumed to be finite). As previously mentioned,

the rate a_n is not necessarily $\sqrt{n}$. Two major avenues resulting in a rate that is less than $\sqrt{n}$ are heavy tailed data, and/or long-range dependence.

To fix ideas, suppose that $a_n = n^\alpha$ for some unknown $\alpha \in (0, 1]$. A successful self-normalization entails constructing a positive statistic, say $\widehat{\zeta}_n > 0$, that converges to some limit distribution Q at a rate explicitly related to the unknown α. For example, consider the case where $n^{\alpha-\delta}\widehat{\zeta}_n$ has limit distribution Q for some known value $\delta > 0$. Then, subsampling can be applied to the self-normalized quantity $(\widehat{\theta}_n - \theta)/\widehat{\zeta}_n$ that converges to a well-defined distribution at the *known* rate n^δ. Strictly speaking, the joint convergence of $(n^\alpha(\widehat{\theta}_n - \theta), n^{\alpha-\delta}\widehat{\zeta}_n)$ to the pair (J, Q) is required; see e.g. Politis et al. (1999, Theorem 11.3.1) for a precise statement.

Typically, the search for a suitable $\widehat{\zeta}_n$ starts with an estimated standard deviation for $\widehat{\theta}_n$—hence the connection to studentization. The self-normalization method for subsampling was first used by Hall et al. (1998) in the context of a long-range dependent data. Hall and Yao (2003) also employ self-normalization in connection with bootstrap with smaller resample size on the residuals of heavy tailed GARCH processes. As is well-known, the bootstrap with smaller resample size is closely related to subsampling in the i.i.d. case; see Politis et al. (1999, Ch. 2.3). The case of self-normalized subsampling for heavy tailed time series was addressed by McElroy and Politis (2002), and Kokoszka and Wolf (2004). McElroy and Politis (2006) is the only paper to-date that achieves a self-normalization in a setting exhibiting both heavy-tails *and* long-range dependence.

References

Ango Nze, P., Bühlmann, P. and Doukhan, P. (2002): Weak dependence beyond mixing and asymptotics for nonparametric regression. *Annals of Statistics* **30**, 397–430.

Ango Nze, P., Dupoiron, S. and Rios, R. (2003): Subsampling under weak dependence conditions. *Tech. Report DT2003-42. CREST, Paris.*

Berkes, I., Horváth, L. and Kokoszka, P. (2003): GARCH processes: structure and estimation. *Bernoulli* **9**, 201–228.

Bertail, P., Politis, D. N. and Romano, J. P. (1999): On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association* **94**, 569–579.

Bollerslev, T. (1986): Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **31**, 307–327.

Bougerol, P. and Picard, N. (1992): Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* **52**, 115–127.

Bühlmann, P. (2002): Bootstrap for time series. *Statistical Science* **17**, 52–72.

Carlstein, E. (1986): The use of subseries values for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics* **14**, 1171–1179.

Carrasco, M. and Chen, X. (2002): Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17–39.

Doukhan, P. and Louhichi, S. (1999): A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and its Applications* **84**, 313–342.

Engle, R. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987–1008.

Fan, J. and Yao, Q. (1998): Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.

Franke, J., Kreiss, J.-P. and Mammen, E. (2002a): Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli* **8**, 1–37.

Franke, J., Kreiss, J.-P., Mammen, E. and Neumann, M. H. (2002b): Properties of the nonparametric autoregressive bootstrap. *Journal of Time Series Analysis* **23**, 555–585.

Franke, J., Neumann, M. H. and Stockis, J. P. (2004): Bootstrapping nonparametric estimators of the volatility function. *Journal of Econometrics* **118**, 189–218.

Giraitis, L., Kokoszka, P. and Leipus, R. (2000): Stationary ARCH models: Dependence structure and central limit theorem. *Econometric Theory* **16**, 3–22.

Hall, P. (1992): *The bootstrap and Edgeworth expansion*. Springer Verlag, New York.

Hall, P., Jing, B.-Y. and Lahiri, S. N. (1998): On the sampling window method for long-range dependent data. *Statistica Sinica* **8**, 1189–1204.

Hall, P. and Yao, Q. (2003): Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71**, 285–317.

Härdle, W., Horowitz, J. and Kreiss, J.-P. (2003): Bootstrap for Time Series. *International Statistical Review* **71**, 435–459.

Hart, J. D. (1995): Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* **6**, 115–142.

Horowitz, J. L. (2003): Bootstrap methods for Markov processes. *Econometrica* **71**, 1049–1082.

Kazakevičius, V. and Leipus, R. (2002): On stationarity in the ARCH(∞) model. *Econometric Theory* **18**, 1–16.

Kokoszka, P., Teyssiére, G. and Zhang, A. (2004): Confidence intervals for the autocorrelations of the squares of GARCH sequences. *In: Bubak, M. et al. (Eds.): ICCS*, 827–834. Springer, Berlin.

Kokoszka, P. and Wolf, M. (2004): Subsampling the mean of heavy-tailed dependent observations. *Journal of Time Series Analysis* **25**, 217–234.

Kreiss, J.-P. (2000): Nonparametric estimation and bootstrap for financial time series. *In: Chan, W. S., Li, W. K. and Tong, H. (Eds.): Statistics and Finance: An Interface.* London, Imperial College Press.

Kreiss, J.-P., Neumann, M. H. and Yao, Q. (2008): Bootstrap tests for simple structures in nonparametric time series regression. *Statistics and Its Interface* to appear.

Künsch, H. R. (1989): The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**, 1217–1241.

Lahiri, S. N. (2003): *Resampling methods for dependent data*. Springer, New York.

Liu, R. and Singh, K. (1992): Moving blocks jackknife and bootstrap capture weak dependence. *In: LePage, R. and Billard, L. (Eds.): Exploring the limits of the bootstrap*, 225–248. Wiley, New York.

Maddala, G. S. and Li, H. (1996): Bootstrap based tests in financial models. *In: Maddala, G. S. and Rao, C. R. (Eds.): Handbook of Statistics* **14**, 463–488. Amsterdam, Elsevier.

Mammen, E. (1992): When does the bootstrap work? Asymptotic results and simulations. *Springer Lecture Notes in Statistics* **77**. Singer Verlag, Heidelberg.

McElroy, T. and Politis, D. N. (2002): Robust inference for the mean in the presence of serial correlation and heavy tailed distributions. *Econometric Theory* **18**, 1019–1039.

McElroy, T. and Politis, D. N. (2006): Self-Normalization for heavy-tailed time series with long memory. *Statistica Sinica* to appear.

Miquel, J. A. and Olave, P. (1999): Bootstrapping forecast intervals in ARCH models. *Test* **8**, 345–364.

Neumann, M. H. and Kreiss, J.-P. (1998): Regression type inference in nonparametric autoregression. *Annals of Statistics* **26**, 1570-1613.

Paparoditis, E. and Politis, D. N. (2000): The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics* **52**, 139–159.

Paparoditis, E. and Politis, D. N. (2001a): The local bootstrap for Markov processes. *Journal of Statistical Planning and Inference* **108**, 301–328.

Paparoditis, E. and Politis, D. N. (2002): A Markovian local resampling scheme for nonparametric estimators in time series analysis. *Econometric Theorey* **17**, 540–566.

Paparoditis, E. and Politis, D. N. (2003): Residual-based block bootstrap for unit root testing. *Econometrica* **71**, 813–855.

Politis, D. N. (2003): The impact of bootstrap methods on time series analysis. *Statistical Science* **18**, 219–230.

Politis, D. N. and Romano, J. P. (1994): Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* **22**, 2031–2050.

Politis, D. N. and Romano, J. P. (1994): The stationary bootstrap. *Journal of the American Statistical Association* **89**, 1303–1313.

Politis, D. N., Romano, J. P. and Wolf, M. (1997): Subsampling for heteroskestastic time series. *Journal of Econometrics* **81**, 281–317.

Politis, D. N., Romano, J. P. and Wolf, M. (1999): *Subsampling*. Springer Verlag, New York.

Politis, D. N., Romano, J. P. and Wolf, M. (2004): Inference for Autocorrelations in the Possible Presence of a Unit Root. *Journal of Time Series Analysis* **25**, 251–263.

Rajarshi, M. B. (1990): Bootstrap in Markov-sequences based on estimates of transition densities. *Annals of the Institute of Statistical Mathematics* **42**, 253–268.

Robinson, P. (1983): Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185–207.

Robinson, P. (2001): Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* **47**, 67–84.

Romano, J. P. and Wolf, M. (2001): Subsampling intervals in autoregressive models with linear time trend. *Econometrica* **69**, 1283–1314.

Ruiz, E. and Pascual, L. (2002): Bootstrapping financial time series. *Journal of Economic Surveys* **16**, 271–300.

Shepard, N. (1996): Statistical aspects of ARCH and stochastic volatility. *In: Cox, D. R. and Hinkley, D. V. (Eds.): Time series models in Econometrics, Finance and other fields*, 1–67. Capman and Hall, London.

Shi, S. G. (1991): Local bootstrap. *Annals of the Institute of Statistical Mathematics*. **43**, 667–676.

Straumann, D. (2005): Estimation in conditionally heteroscedastic time series models. *Lecture Notes in Statistics* **181**. Springer Verlag, Berlin.

Taylor, S. (1986): *Modelling financial time series*. Wiley, New York.

Markov Chain Monte Carlo

Michael Johannes and Nicholas Polson

Abstract This chapter provides an overview of Markov Chain Monte Carlo (MCMC) methods. MCMC methods provide samples from high-dimensional distributions that commonly arise in Bayesian inference problems. We review the theoretical underpinnings used to construct the algorithms, the Metropolis-Hastings algorithm, the Gibbs sampler, Markov Chain convergence, and provide a number of examples in financial econometrics.

1 Introduction

The Bayesian solution to any inference problem is a simple rule: compute the conditional distribution of unobserved variables given observed data. In financial time series settings, the observed data is asset prices, $y = (y_1, ..., y_T)$, and the unobservables are a parameter vector, θ, and latent variables, $x = (x_1, ..., x_T)$, and the inference problem is solved by $p(\theta, x|y)$, the posterior distribution. The latent variables are either unobserved persistent states such as expected returns or volatility or unobserved transient shocks such as price jump times or sizes.

Characterizing the posterior distribution, however, is often difficult. In most settings $p(\theta, x|y)$ is complicated and high-dimensional, implying that standard sampling methods either do not apply or are prohibitively expensive in terms of computing time. Markov Chain Monte Carlo (MCMC) methods provide a simulation based method for sampling from these high-dimensional

Michael Johannes
Graduate School of Business, Columbia University, 3022 Broadway, NY, 10027, e-mail: mj335@columbia.edu

Nicholas Polson
Graduate School of Business, University of Chicago, 5807 S. Woodlawn, Chicago IL 60637, e-mail: ngp@chicagogsb.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_43,

distributions, and are particularly useful for analyzing financial time series models that commonly incorporate latent variables. These samples can be used for estimation, inference, and prediction.

MCMC algorithms generate a Markov chain, $\left\{\theta^{(g)}, x^{(g)}\right\}_{g=1}^{G}$, whose stationary distribution is $p(\theta, x|y)$. To do this, the first step is the Clifford-Hammersley (CH) theorem, which states that a high-dimensional joint distribution, $p(\theta, x|y)$, is completely characterized by a larger number of lower dimensional conditional distributions. Given this characterization, MCMC methods iteratively sample from these lower dimensional conditional distributions using standard sampling methods and the Metropolis-Hastings algorithm. Thus, the key to Bayesian inference is simulation rather than optimization.

The simulations are used to estimate integrals via Monte Carlo that naturally arise in Bayesian inference. Common examples include posterior moments of parameters, $E[\theta|y]$, or state variables, $E[x|y]$, or even expected utility. Monte Carlo estimates are given by

$$\begin{aligned}\widehat{E}(f(\theta, x)|y) &= G^{-1}\sum_{g=1}^{G} f\left(\theta^{(g)}, x^{(g)}\right) \\ &\approx \int f(\theta, x)\, p(\theta, x|y)\, d\theta dx = E(f(\theta, x)|y).\end{aligned}$$

The rest of the chapter is outlined as follows. In Section 2, we explain the components and theoretical foundations of MCMC algorithms. Section 3 provides a few examples from financial econometrics, and Section 4 provides a list of notable references.

2 Overview of MCMC Methods

To develop the foundations of MCMC in the simplest setting, we consider sampling from a bivariate posterior distribution $p(\theta_1, \theta_2|y)$, and suppress the dependence on the data for parsimony. For intuition, it is useful to think of θ_1 as traditional static parameters and θ_2 as latent variables.

2.1 Clifford–Hammersley theorem

The Clifford-Hammersley theorem (CH) proves that the joint distribution, $p(\theta_1, \theta_2)$, is completely determined by the conditional distributions, $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$, under a positivity condition. The positivity condition requires that $p(\theta_1, \theta_2)$, $p(\theta_1)$ and $p(\theta_2)$ have positive mass for all points. These re-

sults are useful in practice because in most cases, $p(\theta_1, \theta_2)$ is only known up to proportionality and cannot be directly sampled. CH implies that the same information can be extracted from the lower-dimensional conditional distributions, breaking "curse of dimensionality" by transforming a higher dimensional problem, sampling from $p(\theta_1, \theta_2)$, into easier problems, sampling from $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$.

The CH theorem is based on the Besag formula: for any pairs (θ_1, θ_2) and $(\theta_1^{'}, \theta_2^{'})$,

$$\frac{p(\theta_1, \theta_2)}{p(\theta_1', \theta_2')} = \frac{p(\theta_1|\theta_2^{'})p(\theta_2|\theta_1)}{p(\theta_1'|\theta_2')\, p(\theta_2'|\theta_1)}. \tag{1}$$

The proof uses the fact that $p(\theta_1, \theta_2) = p(\theta_2|\theta_1)\, p(\theta_1)$, which, when applied to (θ_1, θ_2) and $(\theta_1^{'}, \theta_2^{'})$, implies that

$$p(\theta_1) = \frac{p(\theta_1|\theta_2')\, p(\theta_2')}{p(\theta_2'|\theta_1)}.$$

The general version of CH follows by analogy. Partitioning a vector as $\theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_K)$, then the general CH theorem states that

$$p(\theta_i|\theta_{-i}) \triangleq p(\theta_i|\theta_1, \theta_{1,\ldots,}\theta_{i-1}, \theta_{i+1}, ..., \theta_K),$$

for $i = 1, ..., K$, completely characterizes the joint distribution $p(\theta_1, ..., \theta_K)$.

An important case arises frequently in models with latent variables. Here, the posterior is defined over vectors of static fixed parameters, θ, and latent variables, x. In this case, CH implies that $p(\theta, x|y)$ is completely characterized by $p(\theta|x, y)$ and $p(x|\theta, y)$. The distribution $p(\theta|x, y)$ is the posterior distribution of the parameters, conditional on the observed data and the latent variables. Similarly, $p(x|\theta, y)$ is the smoothing distribution of the latent variables given the parameters.

2.2 Constructing Markov chains

To construct the Markov chains for MCMC with the appropriate limiting distribution, we use direct sampling methods for known distributions and otherwise use indirect sampling methods such as the Metropolis-Hastings algorithm. First, we describe the indirect methods and then explain the Gibbs sampler and hybrid algorithms, which combine aspects of Metropolis-Hastings and direct sampling methods.

2.2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm provides a general approach for sampling from a given target density, $\pi(\theta)$. MH uses an accept-reject approach, drawing a candidate from a distribution $q(\theta)$ that is accepted or rejected based on an acceptance probability. Unlike traditional accept-reject algorithms, which repeatedly sample until acceptance, the MH algorithm samples only once at each iteration. If the candidate is rejected, the algorithm keeps the current value. In this original form of the algorithm, the entire vector θ is update at once. Below, modifications are discussed that update θ is blocks, using the intuition from CH.

Specifically, the MH algorithm repeats the following two steps G times: given $\theta^{(g)}$

$$\begin{aligned}&\text{Step 1. Draw } \theta' \text{ from a proposal distribution, } q(\theta'|\theta^{(g)})\\ &\text{Step 2. Accept } \theta' \text{ with probability } \alpha\left(\theta^{(g)}, \theta'\right),\end{aligned}$$

where

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{(g)})}\frac{q(\theta^{(g)}|\theta')}{q(\theta'|\theta^{(g)})}, 1\right).$$

To implement the accept-reject step, draw a uniform random variable, $U \sim U[0,1]$, and set $\theta^{(g+1)} = \theta'$ if $U < \alpha\left(\theta^{(g)}, \theta'\right)$, leaving $\theta^{(g)}$ unchanged ($\theta^{(g+1)} = \theta^{(g)}$) otherwise. It is important to note that the denominator in the acceptance probability cannot be zero, provided the algorithm is started from a $\pi-$positive point since q is always positive. The MH algorithm only requires that π can be evaluated up to proportionality.

The output of the algorithm, $\left\{\theta^{(g)}\right\}_{g=1}^{\infty}$, is clearly a Markov chain. The key theoretical property is that the Markov chain, under mild regularity, has $\pi(\theta)$ as its limiting distribution. We discuss two important special cases that depend on the choice of q.

Independence MH

One special case draws a candidate *independently* of the previous state, $q(\theta'|\theta^{(g)}) = q(\theta')$. In this independence MH algorithm, the acceptance criterion simplifies to

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{(g)})}\frac{q(\theta^{(g)})}{q(\theta')}, 1\right)$$

Even though θ' is drawn independently of the previous state, the sequence generated is not independent, since α depends on previous draws. The criterion implies a new draw is always accepted if target density ratio, $\pi(\theta')/\pi(\theta^{(g)})$, increases more than the proposal ratio, $q(\theta^{(g)})/q(\theta')$. When

this is not satisfied, an balanced coin is flipped to decide whether or not to accept the proposal.

When using independence MH, it is common to pick the proposal density to closely match certain properties of the target distribution. One common criterion is to ensure that tails of the proposal density are thicker than the tails of the target density. By "blanketing" the target density, it is less likely that the Markov chain will get trapped in a low probability region of the state space.

Random-walk Metropolis

Random-walk (RW) Metropolis is the polar opposite of the independence MH algorithm. It draws a candidate from the following RW model,

$$\theta' = \theta^{(g)} + \sigma\varepsilon_{g+1},$$

where ε_t is an independent, mean zero, and symmetric error term, typically taken to be a normal or $t-$distribution, and σ is a scaling factor. The algorithm must be tuned via the choice of σ, the scaling factor. Symmetry implies that

$$q\left(\theta'|\theta^{(g)}\right) = q\left(\theta^{(g)}|\theta'\right),$$

with acceptance probability

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\pi(\theta')/\pi(\theta^{(g)}), 1\right).$$

The RW algorithm, unlike the independence algorithm, learns about the density $\pi(\theta)$ via small symmetric steps, randomly "walks" around the support of π. If a candidate draw has a higher target density value than the current draw, $\pi(\theta') > \pi(\theta^{(g)})$, the draw is always accepted. If $\pi(\theta') < \pi(\theta^{(g)})$, then a unbalanced coin is flipped.

2.2.2 Gibbs sampling

The Gibbs sampler simulates multidimensional posterior distributions by iteratively sampling from the lower-dimensional conditional posteriors. The Gibbs sampler updates the chain one component at a time, instead of updating the entire vector. This requires either that the conditional posteriors distributions are discrete, are a recognizable distribution (e.g. normal) for which standard sampling algorithms apply, or that resampling methods, such as accept-reject, can be used.

In the case of $p(\theta_1, \theta_2)$, given current draws, $\left(\theta_1^{(g)}, \theta_2^{(g)}\right)$, the Gibbs sampler consists of

$$1.\ \text{Draw } \theta_1^{(g+1)} \sim p\left(\theta_1|\theta_2^{(g)}\right)$$
$$2.\ \text{Draw } \theta_2^{(g+1)} \sim p\left(\theta_2|\theta_1^{(g+1)}\right),$$

repeating G times. The draws generated by the Gibbs sampler form a Markov chain, as the distribution of $\theta^{(g+1)}$ conditional on $\theta^{(g)}$ is independent of past draws. Higher dimensional cases follow by analogy.

2.2.3 Hybrid chains

Given a partition of the vector θ via CH, a hybrid MCMC algorithm updates the chain one subset at a time, either by direct draws ('Gibbs steps') or via MH ('Metropolis step'). Thus, a hybrid algorithm combines the features of the MH algorithm and the Gibbs sampler, providing significant flexibility in designing MCMC algorithms for different models.

To see the mechanics, consider the two-dimensional example. First, assume that the distribution $p(\theta_2|\theta_1)$ is recognizable and can be directly sampled. Second, suppose that $p(\theta_1|\theta_2)$ can only be evaluated and not directly sampled. Thus we use a Metropolis step to update θ_1 given θ_2. For the MH step, the candidate is drawn from $q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)$, which indicates that the step can depend on the past draw for θ_1. We denote the Metropolis step as $MH\left[q\left(\theta_1|\theta_1^{(g)}, \theta_2^{(g)}\right)\right]$, which implies that we draw $\theta_1^{(g+1)} \sim q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)$ and then accept/reject based on

$$\alpha\left(\theta_1^{(g)}, \theta_1'\right) = \min\left(\frac{p\left(\theta_1'|\theta_2^{(g)}\right)}{p\left(\theta_1^{(g)}|\theta_2^{(g)}\right)} \frac{q\left(\theta_1^{(g)}|\theta_1', \theta_2^{(g)}\right)}{q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)}, 1\right).$$

The general hybrid algorithm is as follows. Given $\theta_1^{(g)}$ and $\theta_2^{(g)}$, for $g = 1, ..., G$,

$$1.\ \text{Draw } \theta_1^{(g+1)} \sim MH\left[q\left(\theta_1|\theta_1^{(g)}, \theta_2^{(g)}\right)\right]$$
$$2.\ \text{Draw } \theta_2^{(g+1)} \sim p\left(\theta_2|\theta_1^{(g+1)}\right).$$

In higher dimensional cases, a hybrid algorithm consists of any combination of Gibbs and Metropolis steps. Hybrid algorithms significantly increase the applicability of MCMC methods, as the only requirement is that the model generates posterior conditionals that can either be sampled or evaluated.

2.3 Convergence theory

To understand why MCMC algorithms work, we briefly discuss convergence of the underlying Markov chain for the case of the Gibbs sampler. The arguments for convergence of MH or hybrid algorithms are similar.

The Markov transition kernel from state θ to state θ' is $\mathbb{P}(\theta, \theta') = p(\theta_1'|\theta_2)\, p(\theta_2'|\theta_1')$,and by definition, $\int \mathbb{P}(\theta, \theta')\, d\theta' = 1$. The densities $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ will typically have either discrete or continuous support, and in nearly all cases the chain can reach any point or set in the state space in one step. To establish convergence, we first identify the limiting distribution. A stationary probability distribution, π, satisfies the integral equation

$$\pi(\theta') = \int \mathbb{P}(\theta, \theta')\, \pi(\theta)\, d\theta.$$

If the chain converges, then π is also called the limiting distribution. It is easy to verify that the stationary distribution of the Markov chain generated by the Gibbs sampler is the posterior distribution, $\pi(\theta) = p(\theta_1, \theta_2)$:

$$\begin{aligned}\int \mathbb{P}(\theta, \theta')\, p(\theta)\, d\theta &= p(\theta_2'|\theta_1') \int_{\theta_2} \int_{\theta_1} p(\theta_1'|\theta_2)\, p(\theta_1, \theta_2)\, d\theta_1 d\theta_2 \\ &= p(\theta_2'|\theta_1') \int_{\theta_2} p(\theta_1'|\theta_2)\, p(\theta_2)\, d\theta_2 \\ &= p(\theta_2'|\theta_1')\, p(\theta_1') = p(\theta_1', \theta_2') = \pi(\theta').\end{aligned}$$

To establish convergence to the limiting distribution, the chain must satisfy certain regularity conditions on how it traverses the state space. Starting from an initial π-positive point, the Markov chain in Gibbs samplers can typically reach any set in the state space in one step, implying that states communicate and the chain is irreducible. This does not imply that a chain starting from a given point, will return to that point or visit nearby states *frequently*. Well-behaved chains are not only irreducible, but stable, in the sense that they make many return visits to states. Chains that visit states or sets frequently are recurrent. Under very mild conditions, the Gibbs sampler generates an irreducible and recurrent chain. In most cases, a measure theoretical condition called Harris recurrence is also satisfied, which implies that the chains converge for any starting values.

In this case, the ergodic theorem holds: for a sufficiently integrable function f and for all starting points θ,

$$\lim_{G\to\infty} \frac{1}{G} \sum_{g=1}^{G} f\left(\theta^{(g)}\right) = \int f(\theta)\, \pi(\theta)\, d\theta = E[f(\theta)]$$

almost surely. Notice the two subtle modes of convergence: there is the convergence of the Markov chain to its stationary distribution, and Monte Carlo convergence, which is the convergence of the partial sums to the integral.

In practice, a chain is typically run for an initial length, often called the burn-in, to remove any dependence on the initial conditions. Once the chain has converged, then a secondary sample of size G is created for Monte Carlo inference.

3 Financial Time Series Examples

While there are many examples of MCMC methods analyzing financial time series models, we focus on just three prominent examples, providing references at the end for other applications.

3.1 Geometric Brownian motion

The geometric Brownian motion is the simplest model,

$$y_t = \mu + \sigma\varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ and y_t are continuously compounded returns. The likelihood function is $p\left(y|\mu,\sigma^2\right)$, and $p\left(\mu,\sigma^2|y\right)$ is the joint posterior. We assume independent conjugate priors, $p(\mu) \sim \mathcal{N}(a,A)$ and $p\left(\sigma^2\right) \sim \mathcal{IG}\left(\frac{b}{2},\frac{B}{2}\right)$, where $\mathcal{IG}$ denotes the inverse Gamma distribution, and a, A, b, and B are hyperparameters.

CH implies that $p\left(\mu|\sigma^2,y\right)$ and $p\left(\sigma^2|\mu,y\right)$ are the complete conditionals, which are given by Bayes rule as

$$\begin{aligned} p\left(\mu|\sigma^2,y\right) &\propto p\left(y|\mu,\sigma^2\right)p(\mu) \\ p\left(\sigma^2|\mu,y\right) &\propto p\left(y|\mu,\sigma^2\right)p\left(\sigma^2\right). \end{aligned}$$

Straightforward algebra implies that

$$p\left(\mu|y,\sigma^2\right) \sim \mathcal{N}\left(a^T,A^T\right) \text{ and } p\left(\sigma^2|y,\mu\right) \sim \mathcal{IG}\left(\frac{b^T}{2},\frac{B^T}{2}\right),$$

where

$$\begin{aligned} a^T &= A^T\left(\frac{\overline{y}}{\sigma^2/T}+\frac{a}{A}\right),\ A^T = \left(\frac{1}{\sigma^2/T}+\frac{1}{A}\right)^{-1} \\ b^T &= b+T \text{ and } B^T = B+\sum\nolimits_{t=1}^{T}\left(y_t-\mu\right)^2, \end{aligned}$$

where $T^{-1}\sum_{t=1}^{T} y_t = \overline{y}$.

The fact that the conditional posterior is the same distribution (with different parameters) as the prior distribution is a property of the prior known as conjugacy.

Since both distributions are standard distributions, the MCMC algorithm is a two-step Gibbs sampler. Given current draws, $\left(\mu^{(g)}, \left(\sigma^2\right)^{(g)}\right)$, the algorithm iteratively simulates

$$\begin{aligned}
&1.\ \text{Draw } \mu^{(g+1)} \sim p\left(\mu|\left(\sigma^2\right)^{(g)}, y\right) \sim \mathcal{N} \\
&2.\ \text{Draw } \left(\sigma^2\right)^{(g+1)} \sim p\left(\sigma^2|\mu^{(g+1)}, y\right) \sim \mathcal{IG}.
\end{aligned}$$

This example is meant to develop intuition. In most cases, one would chose a dependent prior of the form

$$p\left(\mu, \sigma^2\right) \propto p\left(\mu|\sigma^2\right) p\left(\sigma^2\right),$$

where $p\left(\mu|\sigma^2\right) \sim \mathcal{N}$ and $p\left(\sigma^2\right) \sim \mathcal{IG}$. This is known as the $\mathcal{NIG}$ is the normal-inverse gamma joint prior. In this case, MCMC is not required as one can draw directly from $p\left(\mu, \sigma^2|y\right)$.

3.2 Time-varying expected returns

Next, consider a model with time-varying expected returns,

$$\begin{aligned}
y_t &= \mu_t + \sigma\varepsilon_t \\
\mu_t &= \alpha_\mu + \beta_\mu \mu_{t-1} + \sigma_\mu \varepsilon_t.
\end{aligned}$$

The parameter vector is $\theta = \left(\sigma^2, \alpha_\mu, \beta_\mu, \sigma_\mu^2\right)$ and the state variables are $\mu = (\mu_1, ..., \mu_T)$. We assume standard conjugate priors, $\sigma^2 \sim \mathcal{IG}$ and $(\alpha_\mu, \beta_\mu, \sigma_\mu) \sim \mathcal{NIG}$, suppressing the parameter of these distributions. CH implies that $p\left(\sigma^2|\alpha_\mu, \beta_\mu, \sigma_\mu^2, \mu, y\right)$, $p\left(\alpha_\mu, \beta_\mu, \sigma_\mu^2|\sigma^2, \mu, y\right)$, and $p\left(\mu|\sigma^2, \alpha_\mu, \beta_\mu, \sigma_\mu^2, y\right)$ are the complete conditionals.

The Gibbs sampler for this model is given by:

$$\begin{aligned}
&1.\ \left(\sigma^2\right)^{(g+1)} \sim p\left(\sigma^2|\alpha_\mu^{(g)}, \beta_\mu^{(g)}, \left(\sigma_\mu^2\right)^{(g)}, \mu^{(g)}, y\right) \sim \mathcal{IG} \\
&2.\ \left(\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}, \left(\sigma_\mu^2\right)^{(g+1)}\right) \sim p\left(\alpha_\mu, \beta_\mu, \sigma_\mu^2|\left(\sigma^2\right)^{(g+1)}, \mu^{(g)}, y\right) \sim \mathcal{NIG} \\
&3.\ \mu^{(g+1)} \sim p\left(\mu|\left(\sigma^2\right)^{(g+1)}, \alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}, \left(\sigma_\mu^2\right)^{(g+1)}, y\right) \sim FFBS,
\end{aligned}$$

where the third step refers to the forward-filtering, backward sampling algorithm. This algorithm applies in conditionally Gaussian state space models,

and requires three steps:

Step 1. Run the Kalman filter forward for $t = 1, ..., T$ to get the moments of $p\left(\mu_t|\theta, y^t\right)$

Step 2. Sample the last state from $\widehat{\mu}_T \sim p\left(\mu_T|\theta, y^T\right)$

Step 3. Sample backward through time: $\widehat{\mu}_t \sim p\left(\mu_t|\widehat{\mu}_{t+1}, \theta, y^t\right).$

where $y^t = (y_1, ..., y_t)$. The FFBS algorithm provides a direct draw of the vector μ from its conditional distribution, which is more efficient than sampling the expected returns, μ_t, one state at a time.

The output of the algorithm can be used for Monte Carlo integration. For example, the smoothed estimate of the latent state at time t is given by

$$\frac{1}{G}\sum_{g=1}^{G}\mu_t^{(g)} \approx \int \mu_t p\left(\mu_t|y\right) d\mu_t = E\left(\mu_t|y\right).$$

3.3 Stochastic volatility models

A popular discrete-time stochastic volatility model is given by

$$\begin{aligned} y_t &= \sqrt{V_{t-1}}\varepsilon_t \\ \log\left(V_t\right) &= \alpha_v + \beta_v \log\left(V_{t-1}\right) + \sigma_v \varepsilon_t^v, \end{aligned}$$

where, for simplicity, we assume the errors are uncorrelated. Again, a $\mathcal{NIG}$ prior for $\left(\alpha_v, \beta_v, \sigma_v^2\right)$ is conjugate for the parameters, conditional on the volatilities.

The only difficulty in this model is sampling from $p\left(V|\alpha_v, \beta_v, \sigma_v^2, y\right)$. This distribution is not a recognizable distribution, and due to its high dimension, a direct application of MH is not recommended. The simplest approach is to use the CH theorem to break the T-dimensional distribution $p\left(V|\alpha_v, \beta_v, \sigma_v^2, y\right)$ into T 1-dimensional distributions,

$$p\left(V_t|V_{t-1}, V_{t+1}, \theta, y_{t+1}\right) \propto p\left(y_{t+1}|V_t\right) p\left(V_{t+1}|V_t, \theta\right) p\left(V_t, |V_{t-1}, \theta\right),$$

for $t = 1, ..., T$. This distribution is again not recognizable, but it is easy to develop proposal distributions that closely approximate the distribution using independence MH, although the random-walk algorithm also applies and works well in practice. This is typically referred to as a single-state volatility updating step.

Thus, the hybrid MCMC algorithm for estimating the stochastic volatility requires the following steps: given

$$1.\ \left(\alpha_v^{(g+1)}, \beta_v^{(g+1)}, \left(\sigma_v^2\right)^{(g+1)}\right) \sim p\left(\alpha_v, \beta_v, \sigma_v^2 | V^{(g)}, y\right) \sim \mathcal{NIG}$$

$$2.\ V_t^{(g+1)} \sim MH\left[q\left(V_t | V_{t-1}^{(g)}, V_t^{(g)}, V_{t+1}^{(g)}, \theta^{(g+1)}\right)\right] \text{ for } t = 1, ..., T.$$

When implementing this model, care needs to be taken with the Metropolis step. It is common to try alternative proposal distribution and perform simulation studies to ensure the algorithm is working properly.

4 Further Reading

For a textbook discussion of the Bayesian approach to inference, we recommend the books by Raiffa and Schlaifer (1961), Bernardo and Smith (1995), Robert (2001), or O'Hagan (2004). Robert and Casella (2005) or Gamerman and Lopes (2006) provide excellent textbook treatments of MCMC methods. They provide with details regarding the algorithms (e.g., tuning MH algorithms) and numerous examples.

It is impossible to cite all of the important papers developing MCMC theory and building MCMC algorithms in different applications. We here provide the briefest possible list, with an emphasis on the initial MCMC approaches for various different models. The extensions to these foundational papers are numerous.

One important precursor to MCMC methods in Bayesian statistics is Tanner and Wong (1987), who introduced algorithms using data augmentation. Gelfand and Smith (1990) provided the first MCMC applications in Bayesian statistics. Smith and Roberts (1993) and Besag, Green, Higdon, and Mengersen (1995) provide overviews of MCMC methods.

Regarding the underlying theory of MCMC algorithms, The Clifford-Hammersley theorem was originally shown in Hammersley and Clifford (1970) and the Besag formula is in Besag (1974). The original Metropolis random-walk algorithm is given in Metropolis et al. (1953), and the independence version in Hastings (1973). Geman and Geman (1984) introduced the Gibbs sampler for sampling posterior distributions and proved convergence properties. Tierney (1994) provides a wide range of theoretical convergence results for MCMC algorithms, providing verifiable conditions for various forms of convergence and discussing hybrid algorithms. Chib and Greenberg (1995) provide an overview of the Metropolis-Hastings algorithm.

With regard to specific models, there are a number of important foundational references. For simplicity, we list them in chronological order. Carlin and Polson (1991) developed MCMC algorithms for models with scale mixture of normal distribution errors, which includes the t, double exponential,

logistic, and exponential power family error distributions. Carlin and Polson (1992) and develop MCMC algorithms for discrete regression and categorical observations and for the probit model, see Albert and Chib (1993). Carlin, Gelfand, and Smith (1992) and Chib (1998) developed algorithms for time series models with change-points. Diebold and Robert (1994) analyzed finite mixture models with MCMC methods. Carlin, Polson, and Stoffer (1992) develop MCMC methods for nonlinear and non-normal state space models, and Carter and Kohn (1994, 1996) developed the FFBS algorithm for estimation in a range of non-normal state space models. McCulloch and Tsay (1993) analyze Markov switching models.

MCMC methods have been broadly applied in stochastic volatility models. Jacquier, Polson, and Rossi (1994) first developed MCMC algorithms for the log-stochastic volatility models, with Jacquier, Polson, and Rossi (2004) providing extensions to correlated and non-normal error distributions. Eraker, Johannes, and Polson (2003) analyzed time series models with jumps in prices and volatility. Jones (1998), Eraker (2001) and Elerian, Shephard, and Chib (2001) develop approaches for MCMC analysis of continuous-time models by simulating additional high-frequency data points between observations. Also, see the chapter by Chib, Omori, and Asai in this handbook for further references for multivariate problems. For a more extensive review of MCMC methods for financial econometrics, see Johannes and Polson (2005).

References

Albert, J. and Chib, S. (1993): Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88**, 669–679.

Bernardo, J. and Smith, A. (1995): *Bayesian Theory* Wiley, New York.

Besag, J. (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36**, 192–326.

Besag, J. Green, E., Higdon, D. and Mengersen, K. (1995): Bayesian computuation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Carlin, B. and Polson, N. (1991): Inference for Nonconjugate Bayesian Models using the Gibbs sampler. *Canadian Journal of Statistics* **19**, 399–405.

Carlin, B. and Polson, N. (1992): Monte Carlo Bayesian Methods for Discrete Regression Models and Categorical Time Series. *In: Bernardo, J.M. et al (Eds.): Bayesian Statistics* **4**, 577–586. Oxford University Press, Oxford.

Carlin, B., Polson, N. and Stoffer, D. (1992): A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association* **87**, 493–500.

Carlin, B., Gelfand, A. and Smith, A. (1992): Hierarchical Bayesian analysis of change point process. *Applied Statistics, Series C* **41**, 389–405.

Carter, C., and Kohn, R. (1994): On Gibbs Sampling for State Space Models. *Biometrika* **81**, 541–553.

Carter, C. and Kohn,R. (1996): Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika* **83**, 589–601.

Chib, S. (1998): Estimation and Comparison of Multiple Change Point Models. *Journal of Econometrics* **86**, 221–241.

Chib, S. and Greenberg, E. (1995): Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49**, 327–335.

Diebolt, J. and Robert, C. (1994): Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical Society Series B* **56**, 363–375.

Elerian, O., Shephard, N. and Chib, S. (2001): Likelihood inference for discretely observed non-linear diffusions. *Econometrica* **69**, 959–993.

Eraker, B. (2001): MCMC Analysis of Diffusion Models with Applications to Finance. *Journal of Business and Economic Statistics* **19**, 177–191.

Eraker, B., Johannes, M. and Polson, N. (2003): The Impact of Jumps in Equity Index Volatility and Returns. *Journal of Finance* **58**, 1269–1300.

Gamerman, D. and Lopes, H. (2006): *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* Boca Raton, Chapman & Hall/CRC.

Gelfand, A. and Smith, A. (1990): Sampling Based approaches to calculating Marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Geman, S. and Geman, D. (1984): Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hammersley, J. and Clifford, P. (1970): *Markov fields on finite graphs and lattices. Unpublished Manuscipt.*

Hastings, W. K. (1970): Monte Carlo sampling Methods using Markov Chains and their Applications. *Biometrika* **57**, 97–109.

Jacquier, E., Polson, N. and Rossi, P. (1994): Bayesian analysis of Stochastic Volatility Models (with discussion). *Journal of Business and Economic Statistics* **12**, 371–417.

Jacquier, E., Polson, N. and Rossi, P. (2004): Bayesian Inference for SV models with Correlated Errors. *Journal of Econometrics* forthcoming.

Johannes, M. and Polson, N. (2005): MCMC methods for Financial Econometrics. *In: Ait-Sahalia, Y. and Hansen, L. (Eds.): Handbook of Financial Econometrics* forthcoming.

Jones, C. (1998): Bayesian Estimation of Continuous-Time Finance Models. *Working paper.*

McCulloch, R. and Tsay, R. (1993): Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series. *Journal of the American Statistical Association* **88**, 968–978.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953): Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21**, 1087–1091.

O'Hagan, A. and Forster, J. (2004): *Kendall's Advanced Theory of Statistics: Bayesian Inference.* **2**, Hodder-Arnold.

Raiffa, H. and Schlaifer, R. (1961): *Applied Statistical Decision Theory* Harvard University, Boston, MA.

Robert, C. (2001): *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, Springer-Verlag, New York.

Robert, C. and Casella, G. (2005): *Monte Carlo Statistical Methods* New York, Springer.

Smith, A. and Gareth, R. (1993): Bayesian Computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Association Series B* **55**, 3–23.

Tanner, M. and Wong, W. (1987): The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Tierney, L. (1994): Markov Chains for exploring Posterior Distributions (with discussion). *Annals of Statistics* **22**, 1701–1786.

Particle Filtering

Michael Johannes and Nicholas Polson

Abstract This chapter provides an overview of particle filters. Particle filters generate approximations to filtering distributions and are commonly used in non-linear and/or non-Gaussian state space models. We discuss general concepts associated with particle filtering, provide an overview of the main particle filtering algorithms, and provide an empirical example of filtering volatility from noisy asset price data.

1 Introduction

Filtering generally refers to an extraction process, and statistical filtering refers to an algorithm for extracting a latent state variable from noisy data using a statistical model. The original filtering applications were in physical systems, for example, tracking the location of an airplane or missile using noisy radar signals, but filtering quickly became a crucial theoretical and empirical tool in economics and finance due to widespread use of models incorporating latent variables.

Latent variables capture unobserved changes in the economic environment. In many cases, there is clear evidence for time-variation, but the underlying causes are unknown or difficult to quantify. For example, in finance, it is clear that asset return volatility time-varies, but it is difficult to identify factors generating the variation. To capture the variation, common models assume that volatility is unobserved but evolves through time as a stochastic

Michael Johannes
Graduate School of Business, Columbia University, 3022 Broadway, NY, 10027, e-mail: mj335@columbia.edu

Nicholas Polson
Graduate School of Business, University of Chicago, 5807 S. Woodlawn, Chicago IL 60637, e-mail: ngp@chicagogsb.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_44,

process. In macroeconomics, a similar challenge occurs when modeling time-varying components that drive aggregates. For example, the key component in "long-run risk" models is a highly persistent unobserved variable driving consumption and dividend growth rates.

Given the widespread use of latent variables in economic models, a central challenge is to develop statistical methods for estimating the latent variables. This chapter discusses a computational approach for filtering known as the particle filter.

To understand the nature of the filtering problem and why particle filters are useful, the formal problem is defined as follows. A statistical model generates the observed data, a vector y_t, and the conditional distribution of y_t depends on a latent state variable, x_t. Formally, the data is generated by the state space model, which consists of the observation and state evolution equations,

$$\begin{aligned} \text{Observation equation: } y_t &= f\left(x_t, \varepsilon_t^y\right) \\ \text{State evolution: } x_{t+1} &= g\left(x_t, \varepsilon_{t+1}^x\right), \end{aligned}$$

where ε_{t+1}^y is the observation error or "noise," and ε_{t+1}^x are state shocks. The observation equation is often written as a conditional likelihood, $p\left(y_t|x_t\right)$, and the state evolution as $p\left(x_{t+1}|x_t\right)$. Both of these distributions typically depend on static parameters, θ, whose dependence is suppressed, except where explicitly noted.

The posterior distribution of x_t given the observed data, $p\left(x_t|y^t\right)$, solves the filtering problem, where $y^t = (y_1, ..., y_t)$ is the observed data. Beginning with Kalman's filter, computing $p\left(x_t|y^t\right)$ uses a two-step procedure of prediction and Bayesian updating. The prediction step combines the current filtering distribution with the state evolution,

$$p\left(x_{t+1}|y^t\right) = \int p\left(x_{t+1}|x_t\right) p\left(x_t|y^t\right) dx_t, \tag{1}$$

providing a forecast of next period's state. Next, given a new observation, y_{t+1}, the predictive or "prior" views are updated by Bayes rule

$$\underbrace{p\left(x_{t+1}|y^{t+1}\right)}_{\text{Posterior}} \propto \underbrace{p\left(y_{t+1}|x_{t+1}\right)}_{\text{Likelihood}} \underbrace{p\left(x_{t+1}|y^t\right)}_{\text{Prior}}. \tag{2}$$

The problem is that $p\left(x_t|y^t\right)$ is known analytically only in a limited number of settings, such as a linear, Gaussian model where $p\left(y_t|x_t\right) \sim \mathcal{N}\left(x_t, \sigma^2\right)$ and $p\left(x_{t+1}|x_t\right) \sim \mathcal{N}\left(x_t, \sigma_x^2\right)$. In this case, the Kalman filter implies that $p(x_t|y^t) \sim N\left(\mu_t, \sigma_t^2\right)$, where μ_t and σ_t^2 solve the Kalman recursions. In nonlinear or non-normal models, it is not possible to analytically compute $p(x_t|y^t)$. In these settings, $p\left(x_t|y^t\right)$ is a complicated function of y^t, and simulation methods are typically required to characterize $p\left(x_t|y^t\right)$.

The particle filter is the most popular approach. A particle filter simulates approximate samples from $p(x_t|y^t)$, which are used for Monte Carlo integration to estimates moments of interest such as $E[f(x_t)|y^t]$. Particle filters use a discrete approximation to $p(x_t|y^t)$ consisting of states or "particles", $\left\{x_t^{(i)}\right\}_{i=1}^N$, and weights associated with those particles, $\left\{\pi_t^{(i)}\right\}_{i=1}^N$. A particle approximation is just a random histogram.

Recursive sampling is the central challenge in particle filtering: given a sample from $p^N(x_t|y^t)$, how to generate a random sample from the particle approximation to $p(x_{t+1}|y^{t+1})$ after receiving a new data point y_{t+1}? Essentially this is a problem of using a discrete approximation to the integral in (1) and then sampling from $p^N\left(x_{t+1}|y^{t+1}\right)$. Since there are multiple ways to sample from a given distribution, there are many different particle filters. For example, importance sampling is commonly used to sample from non-standard distributions, and different importance densities generate different particle filters.

This chapter provides an introduction to these particle filters, and outlines a number of the most common algorithms. Before delving into details, it is important to understand the two main reasons why particle filters are so popular. First, particle filters are very flexible and adaptable. Like all Monte Carlo methods, particle filters can be adapted to the particular model specification under consideration. In particular, it is possible to develop accurate filters for non-linear models with fat-tailed and asymmetric error distributions. These are particularly important for applications where errors are often not normally distributed. Second, particle filters are easy to program and computationally very fast to run, in terms of computing time. For these reasons, particle filters provide an attractive filtering methodology.

2 A Motivating Example

The following provides a common setting in which particle filters are useful. Consider a simple log-stochastic volatility model

$$\begin{aligned} y_{t+1} &= \sqrt{V_{t+1}}\varepsilon_{t+1}^y \\ \log(V_{t+1}) &= \alpha_v + \beta_v \log(V_t) + \sigma_v \varepsilon_{t+1}^v, \end{aligned}$$

where $\varepsilon_{t+1}^y$and ε_{t+1}^y are independent standard normal and V_t is the conditional variance. Again, the parameters are assumed to be known. This is a benchmark specification for modeling time-varying volatility. The top panel of Figure 1 provides a simulated sample path of returns from the specification with $\alpha_v = 0$, $\beta_v = 0.95$, and $\sigma_v = 0.10$. By merely observing the data, it is clear that the conditional volatility time varies, as the amplitude of the fluctuations vary over time.

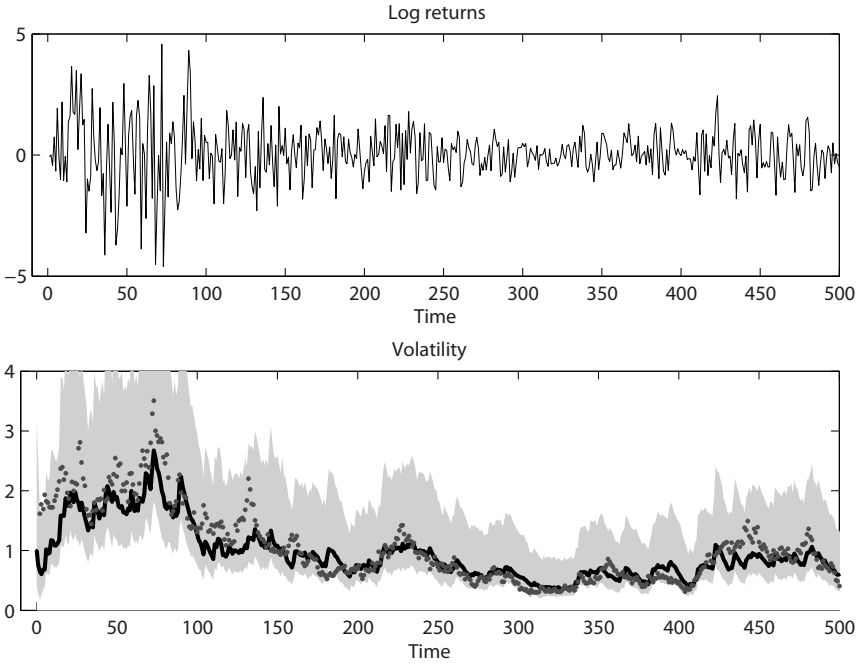

Fig. 1 Simulated returns (top panel) and summaries of the filtering distribution through time (bottom panel) for a logarithmic stochastic volatility model.

This model has conditionally normal errors, but is nonlinear, because of the term $\sqrt{V_{t+1}}\varepsilon_{t+1}^{y}$. Alternatively, the observation equation can be written as

$$\log\left(y_{t+1}\right)^{2} = \log\left(V_{t+1}\right) + \log\left(\varepsilon_{t+1}^{y}\right)^{2},$$

which leads to a model that is linear in $log\left(V_{t+1}\right)$, but is now non-normal because $\log\left(\varepsilon_{t+1}^{y}\right)^{2}$ is $\mathcal{X}^{2}$ with one degree of freedom.

Given this structure, how can V_t be filtered from the observed data? One approach would be to ignore the nonlinearity/non-normality and use the Kalman filter. Alternatively, the model could be linearized using the extended Kalman filter. Both of these approaches are inefficient and biased. Another alternative would be to use deterministic numerical integration to compute the integral in equation (1) and characterize $p\left(V_t|y^t\right)$. This is computationally more difficult, but may work satisfactorily in some settings. In higher dimensional settings, these deterministic numerical integration schemes suffer from Bellman's curse of dimensionality.

Alternatively, the particle filter, as described in the following sections, can be used. The results of the particle filter are displayed in the bottom panel of Figure 1. Here, the true states are given by the dotted line, filtered estimates,

$E\left(\sqrt{V_t}|y^t\right)$, by the solid line, and estimates of the $(5\%, 95\%)$ quantiles in the shades. These are all naturally computed as output from the particle filter.

3 Particle Filters

A particle filter is a discrete approximation, $p^N\left(x_t|y^t\right)$, to $p\left(x_t|y^t\right)$, generally written as $\left\{\pi_t^{(i)}, x_t^{(i)}\right\}_{i=1}^N$, where the weights sum to one, $\sum_{i=1}^N \pi_t^{(i)} = 1$. The support of the discrete approximation, the $x_t^{(i)}$'s, is not preset as would be the case in deterministic approximation schemes, but rather is the stochastic outcome of a simulation algorithm. Thus, the support of the distribution, the $x_t^{(i)}$'s, change from period to period. Thus, a generic particle approximation is given by

$$p^N\left(x_t|y^t\right) = \sum\nolimits_{i=1}^N \pi_t^{(i)} \delta_{x_t^{(i)}},$$

where δ is the Dirac function.

The particle approximation can be transformed into an equally weighted random sample from $p^N\left(x_t|y^t\right)$ by sampling, with replacement, from the discrete distribution, $\left\{\pi_t^{(i)}, x_t^{(i)}\right\}_{i=1}^N$. This procedure, called *resampling*, produces a new sample with uniformly distributed weights, $\pi_t^{(i)} = 1/N$. Resampling can be done in many ways, but the simplest is multinomial sampling. Other methods include stratified sampling, residual resampling, and systematic resampling.

The discrete support of particle filters makes numerical integration easy because integrals becomes sums. For example,

$$p^N\left(x_{t+1}|y^{t+1}\right) \propto \int p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|x_t\right) p^N\left(x_t|y^t\right) dx_t \tag{3}$$

$$\propto \sum_{i=1}^N p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|x_t^{(i)}\right) \pi_t^{(i)}, \tag{4}$$

where the proportionality sign, '$\propto$,' signifies that the normalizing constant does not depend on x_{t+1}. Given the discretization, the central problem in particle filtering is how to generate a sample from $p^N\left(x_{t+1}|y^{t+1}\right)$. Equation (3) implies that $p^N\left(x_{t+1}|y^{t+1}\right)$ is a finite mixture distribution, and different sampling methods generate alternative particle filtering algorithms, each with their own strengths and weaknesses.

In general, there are two sources of approximation errors in particle filtering algorithms. Approximating $p\left(x_t|y^t\right)$ by $p^N\left(x_t|y^t\right)$ generates the first source of error. This is inherent in all particle filtering algorithms, but can be mitigated by choosing N large. Importance sampling or other approximate

sampling methods generate the other source of error, which is present in some particle filtering algorithms. Importance sampling generates an approximate sample from $^{N}(x_t|y^t)$, which in turn approximates $p(x_t|y^t)$. This leads to a second layer of approximation errors.

It is useful to briefly review common uses of the output of particle filtering algorithms. The main use is to estimate latent states. This is done via Monte Carlo. A particle approximation of $E(f(x_t)|y^t)$ is

$$E^N\left(f(x_t)|y^t\right) = \int f(x_t)\, p^N\left(x_t|y^t\right) dx_t = \sum_{i=1}^{N} f\left(x_t^{(i)}\right)\pi_t^{(i)}.$$

As N becomes large, the particle estimates converge by the law of large numbers and a central limit theorem is typically available, both using standard Monte Carlo convergence results.

The filtering distribution is useful for a likelihood based parameter estimation and model comparison. Although in the rest of the chapter we assume parameters are known, a central problem in state space models is estimating the parameters, θ. In the case when parameters are unknown, the particle filter can be used to compute the likelihood function. The likelihood of the observed sample is denoted as $\mathcal{L}\left(\theta|y^T\right)$. In time series models, the likelihood is given by

$$\mathcal{L}\left(\theta|y^T\right) = \prod_{t=0}^{T-1} p\left(y_{t+1}|\theta, y^t\right).$$

In latent variable models, the predictive distribution of the data $p(y_{t+1}|y^t,\theta)$ is not generally known, but rather given as an integral against the filtering distribution:

$$p\left(y_{t+1}|y^t,\theta\right) = \int p(y_{t+1}|\theta,x_{t+1})\, p(x_{t+1}|\theta,x_t)\, p\left(x_t|\theta,y^t\right) dx_t dx_{t+1},$$

where $p(y_{t+1}|\theta,x_{t+1})$ is the conditional likelihood, $p(x_{t+1}|\theta,x_t)$ is the state evolution, and $p(x_t|\theta,y^t)$ is the filtering distribution, all conditional on the unknown parameters. Given a particle approximation to $p(x_t|y^t,\theta)$, it is straightforward to approximate the predictive likelihoods, and therefore to estimate parameters or compare models with likelihood ratios. For the rest of the chapter, we suppress the dependence on the parameters.

The rest of the chapter discusses three common prominent particle filtering algorithms. For parsimony, focus is restricted to particle methods for approximating the filtering distribution, $p(x_t|y^t)$, and we do not discuss methods such as sequential importance sampling (SIS), that generate samples sequentially from the smoothing distribution, $p(x^t|y^t)$, where $x^t = (x_1, ..., x_t)$.

3.1 Exact particle filtering

The easiest way to understand particle filtering is to consider situations in which importance sampling is not required, because direct i.i.d. sampling from $p^N\left(x_{t+1}|y^{t+1}\right)$ is feasible. This is called exact sampling, and leads to an exact particle filter. To see how this works, first note that

$$p\left(y_{t+1}, x_{t+1}|x_t\right) \propto p\left(y_{t+1}|x_t\right) p\left(x_{t+1}|x_t, y_{t+1}\right), \tag{5}$$

which implies that the filtering recursion can be expressed as

$$p\left(x_{t+1}|y^{t+1}\right) \propto \int p\left(y_{t+1}|x_t\right) p\left(x_{t+1}|x_t, y_{t+1}\right) p\left(x_t|y^t\right) dx_t, \tag{6}$$

where $p\left(y_{t+1}|x_t\right)$ is the predictive likelihood and $p\left(x_{t+1}|x_t, y_{t+1}\right)$ is the posterior distribution of the new state given the previous state and the new observation.

This representation generates a different mixture distribution for $p^N\left(x_{t+1}|\right.$ $\left.y^{t+1}\right)$ than the one commonly used in particle filtering algorithms, which is given in 3. Given a particle approximation to $p^N\left(x_t|y^t\right)$,

$$\begin{aligned} p^N\left(x_{t+1}|y^{t+1}\right) &\propto \sum_{i=1}^{N} p\left(y_{t+1}|x_t^{(i)}\right) p\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right) \\ &= \sum_{i=1}^{N} w_t^{(i)} p\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right), \end{aligned} \tag{7}$$

where the normalized first stage weights are

$$w_t^{(i)} = \frac{p\left(y_{t+1}|x_t^{(i)}\right)}{\sum_{i=1}^{N} p\left(y_{t+1}|x_t^{(i)}\right)}.$$

Since $p\left(y_{t+1}|x_t\right)$ is a function of only x_t and y_{t+1}, these weights are known upon receipt of the new observation, which implies that $p^N\left(x_{t+1}|y^{t+1}\right)$ is a standard discrete mixture distribution. Sampling from a discrete mixture distribution is straightforward by first selecting the mixture index and then simulating from that mixture component. This simple procedure leads to exact particle filtering algorithm is

Step 1. Draw $z^{(i)} \sim Mult_N\left(\left\{w_t^{(i)}\right\}_{i=1}^{N}\right)$ for $i = 1, ..., N$ and set $x_t^{(i)} = x_t^{z(i)}$

Step 2. Draw $x_{t+1}^{(i)} \sim p\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right)$ for $i = 1, ..., N$,

where $Mult_N$ denotes an $N-$component multinomial distribution. Since this generates an i.i.d. sample, the second stage weights in the particle approximation $\pi_t^{(i)}$ are all $1/N$.

The intuition of the algorithm is instructive. At Step 1, upon observation of y_{t+1}, the resampling step selects the particles that were most likely, in terms of the predictive likelihood, $p\left(y_{t+1}|x_t^{(i)}\right)$, to have generated y_{t+1}. After this selection step, the algorithm simulates new particles from the component distribution $p\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right)$. The advantage of this algorithm is that it eliminates the second source of errors that can arise in particle filters. By directly sampling from $p^N\left(x_t|y^t\right)$, there are no importance sampling errors. Any remaining Monte Carlo errors can be minimized by choosing N large. It is also possible to sample the discrete distribution in other ways, such as residual or systematic sampling.

The exact particle filtering approach requires that the predictive likelihood

$$p\left(y_{t+1}|x_t\right) = \int p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|x_t\right) dx_{t+1}$$

can be computed and that

$$p\left(x_{t+1}|x_t, y_{t+1}\right) \propto p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|x_t\right)$$

can be sampled. In many models, these distributions are known or straightforward modifications of the general algorithm can be used to generate exact samples from p^N. One such example is given below. In general, exact particle filtering is possible in models with (a) linear observation equations, (b) non-Gaussian errors that can be represented as a discrete or scale mixture of normal distribution, and (c) models with state evolutions that have additive errors, but nonlinear conditional means. This would occur when

$$x_{t+1} = f\left(x_t\right) + \varepsilon_{t+1}^x$$

and $f\left(x_t\right)$ is a known analytical function of x_t. In particular, this allows for a wide range of observation or state errors, including finite mixtures of normal error distributions, t-distributed errors, or double exponential errors. Thus, the class of models in which exact sampling is possible is quite broad.

3.1.1 Example: Linear, Gaussian filtering

To see how exact sampling operates, consider the simple case of filtering in linear Gaussian models: $p\left(y_t|x_t\right) \sim \mathcal{N}\left(x_t, \sigma^2\right)$ and $p\left(x_{t+1}|x_t\right) \sim \mathcal{N}\left(x_t, \sigma_x^2\right)$. The exact or optimal filtering algorithm requires two distributions, $p\left(y_{t+1}|x_t\right)$ and $p\left(x_{t+1}|x_t, y_{t+1}\right)$, which are both easy to characterize:

$$p\left(y_{t+1}|x_t\right) \sim \mathcal{N}\left(x_t, \sigma^2 + \sigma_x^2\right) \text{ and}$$
$$p\left(x_{t+1}|x_t, y_{t+1}\right) \propto p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|x_t\right) \sim \mathcal{N}\left(\mu_{t+1}, \sigma_{t+1}^2\right),$$

where

$$\frac{\mu_{t+1}}{\sigma_{t+1}^2} = \frac{y_{t+1}}{\sigma_y^2} + \frac{x_t}{\sigma_x^2} \text{ and } \frac{1}{\sigma_{t+1}^2} = \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2}.$$

An exact particle filtering algorithm is

$$\text{Step 1: Sample } z^{(i)} \sim Mult_N\left(\left\{w\left(x_t^{(i)}\right)\right\}_{i=1}^N\right)$$
$$\text{and set } x_t^{(i)} = x_t^{z(i)} \text{ for } i = 1, ..., N$$
$$\text{Step 2: Draw } x_{t+1}^{(i)} \sim \mathcal{N}\left(\mu_{t+1}^{(i)}, \sigma_{t+1}^2\right) \text{ for } i = 1, ..., N$$

where

$$w\left(x_t^{(i)}\right) = \exp\left(-\frac{\left(y_{t+1} - x_t^{(i)}\right)^2}{2\left(\sigma^2 + \sigma_x^2\right)}\right) \Bigg/ \sum_{i=1}^N \exp\left(-\frac{\left(y_{t+1} - x_t^{(i)}\right)^2}{2\left(\sigma^2 + \sigma_x^2\right)}\right)$$

and $\mu_{t+1}^{(i)}$ displays the dependence on $x_t^{(i)}$. This generates an equally weighted sample $\left\{x_{t+1}^{(i)}\right\}_{i=1}^N$ from $p^N\left(x_{t+1}|y^{t+1}\right)$, thus $\pi_{t+1}^{(i)} = 1/N$.

3.1.2 Example: Log-stochastic volatility model

A more interesting example is the log-stochastic volatility model, as described in Section 2. The model can be written as

$$\log\left(y_{t+1}\right)^2 = x_{t+1} + \varepsilon_{t+1}$$
$$x_{t+1} = \alpha_v + \beta_v x_t + \sigma_v \varepsilon_{t+1}^v,$$

where ε_{t+1} has a $log\left(\chi_1^2\right)$ distribution. To develop the particle filtering algorithm, it is useful to approximate the $log\left(\chi_1^2\right)$ distribution with a discrete mixture of normals with fixed weights, $\sum_{j=1}^K p_j Z_{t+1}^j$ where $Z_{t+1}^j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ and μ_j and σ_j are known. This approximation is can be made arbitrarily accurate, and in practice 10 mixture components is sufficient.

The key to an efficient particle filter is the introduction of an auxiliary indicator variable, s_{t+1}, that tracks the mixture components. For example, if $s_{t+1} = j$, then

$$p\left(\log\left(y_{t+1}\right)^2 | x_{t+1}, s_{t+1} = j\right) = \mathcal{N}(x_{t+1} + \mu_j, \sigma_j^2).$$

The introduction of an additional latent variable is called data augmentation and is commonly used when developing Markov Chain Monte Carlo algorithms. Given the additional auxiliary variable, the exact sampling algorithm consists of two steps: (1) resampling persistent state variables, in this case, $x_t = log\,(V_t)$, and (2) simulating x_{t+1} and s_{t+1}.

For the first step, the predictive density is

$$p(y_{t+1}|x_t) = \sum_{j=1}^{K} p_j \mathcal{N}(\alpha_v + \beta_v x_t + \mu_j, \sigma_j^2 + \sigma_v^2),$$

and thus

$$w\,(x_t) = \frac{p\,(y_{t+1}|x_t)}{\sum_{i=1}^{N} p\,(y_{t+1}|x_t)}.$$

Note that s_{t+1} is i.i.d., so there is no information available at time t to forecast its value. The second step requires drawing from

$$p\,(x_{t+1}, s_{t+1}|x_t, y_{t+1}) \propto p\,(x_{t+1}|s_{t+1}, x_t, y_{t+1})\,p\,(s_{t+1}|x_t, y_{t+1})\,.$$

The distribution $p\,(s_{t+1}|x_t, y_{t+1})$ is a known discrete distribution since

$$p\,(s_{t+1} = j|x_t, y_{t+1}) \propto p\,(y_{t+1}|x_t, s_{t+1} = j)\,p_j,$$

where

$$p\,(y_{t+1}|x_t, s_{t+1} = j) = \mathcal{N}\left(\alpha_v + \beta_v x_t + \mu_j, \sigma_v^2 + \sigma_j^2\right).$$

Similarly,

$$p\,(x_{t+1}|s_{t+1}, x_t, y_{t+1}) \propto p\,(y_{t+1}|s_{t+1}, x_t)\,p\,(x_{t+1}|x_t)$$

is a convolution of two normal distributions, which is also a normal distribution. Together, these two steps can be used to provide an exact sample from $p^N\left(s_{t+1}, x_{t+1}|y^{t+1}\right)$.

Figure 1 displays a summary of the output of the algorithm, for the simulated path of returns discussed earlier. The bottom panel displays the true simulated volatilities, $\sqrt{V_t}$, in red dots, the posterior mean, $E^N\left(\sqrt{V_t}|y^t\right)$, is the solid line, and the shaded area displays the $(5\%, 95\%)$ quantiles of $p^N\left(\sqrt{V_t}|y^t\right)$.

3.2 SIR

In settings in which exact sampling is not possible, importance sampling is typically used. One of the first, most popular, and most general particle filtering algorithm is known as the sampling importance resampling (SIR) algorithm. The algorithm is simplicity itself, relying only on two steps: given

samples from $p^N(x_t|y^t)$,

$$\text{Step 1. Draw } x_{t+1}^{(i)} \sim p\left(x_{t+1}|x_t^{(i)}\right) \text{ for } i = 1, ..., N$$

$$\text{Step 2. Draw } z^{(i)} \sim Mult_N\left(\left\{w_{t+1}^{(i)}\right\}_{i=1}^N\right) \text{ and set } x_{t+1}^{(i)} = x_{t+1}^{z^{(i)}},$$

where the importance sampling weights are given by

$$w_{t+1}^{(i)} = \frac{p(y_{t+1}|x_{t+1}^{(i)})}{\sum_{i=1}^N p(y_{t+1}|x_{t+1}^{(i)})}.$$

Prior to resampling, each particle had weight $w_{t+1}^{(i)}$. After resampling, the weights are equal, by the definition of resampling. The SIR algorithm has only two mild requirements: that the likelihood function can be evaluated and that the states can be simulated. Virtually every model used in practice satisfies these mild assumptions.

The justification for the algorithm is the weighted bootstrap algorithm or SIR algorithm, which was first developed to simulate posterior distributions, of the form $L(x)\,p(x)$, where L is the likelihood and p the prior. The algorithm first draws an independent sample $x^{(i)} \sim p(x)$ for $i = 1, ...N$, and then computes normalized importance weights $w^{(i)} = L\left(x^{(i)}\right) / \sum_{i=1}^N L\left(x^{(i)}\right)$. The sample drawn from the discrete distribution $\left\{x^{(i)}, w^{(i)}\right\}_{i=1}^N$ tends in distribution to a sample from the product density $L(x)\,p(x)$ as N increases.

In the case of the particle filter, the target density is

$$p\left(x_{t+1}|y^{t+1}\right) \propto p\left(y_{t+1}|x_{t+1}\right) p\left(x_{t+1}|y^t\right). \tag{8}$$

Given an independent sample from $p^N(x_t|y^t)$, the algorithm samples from

$$p^N\left(x_{t+1}|y^t\right) = \int p\left(x_{t+1}|x_t\right) p^N\left(x_t|y^t\right) dx_t,$$

by drawing $x_{t+1}^{(i)} \sim p\left(x_{t+1}|x_t^{(i)}\right)$ for $i = 1, ..., N$. Since $p^N(x_t|y^t)$ is a discrete distribution, this implies that $\left\{x_{t+1}^{(i)}\right\}_{i=1}^N$ is an independent sample from $p^N(x_{t+1}|y^t)$. Resampling with the appropriate weights provides an approximate sample from $p\left(x_{t+1}|y^{t+1}\right)$.

To see the simplicity of the algorithm, consider the benchmark case of filtering in the linear Gaussian model considered earlier. The SIR algorithm requires simulating from $p\left(x_{t+1}|x_t\right) \sim \mathcal{N}\left(x_t, \sigma_x^2\right)$ and evaluating unnormalized weights, which take the form

$$p\left(y_{t+1}|x_{t+1}\right) \propto \exp\left(-\frac{1}{2}\frac{\left(y_{t+1} - x_{t+1}\right)^2}{\sigma^2}\right).$$

3.2.1 Problems with the SIR algorithm

There are a number of problems with SIR. One problem is sample impoverishment or weight degeneracy, which occurs when a vast majority of the weight is placed on a single particle. When this occurs, the resampling step results in a single particle being sampled multiple times. Thus resampling does not fix the sample impoverishment/weight degeneracy problem, it just hides it. Another related problem is that the states are drawn from the prior distribution, $p(x_{t+1}|x_t)$, without accounting for the next period's observation, y_{t+1}. This implies that the simulated states may not be in important or high likelihood, $p(y_{t+1}|x_{t+1})$, regions. In models with outliers, a large y_{t+1} is observed, but the SIR algorithm draws samples from $p\left(x_{t+1}|x_t^{(i)}\right)$ ignoring the new observation.

To mitigate this problem, it is possible to choose an alternative importance density. Instead of drawing from $p(x_{t+1}|x_t)$, it is possible to draw from an importance density that depends on y_{t+1},

$$x_{t+1}^{(i)} \sim q\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right).$$

In this case the unnormalized weights are

$$w_{t+1}^{(i)} \propto \frac{p(y_{t+1}|x_{t+1}^{(i)})p\left(x_{t+1}^{(i)}|x_t^{(i)}\right)}{q\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right)}.$$

The "optimal" importance sampling density, in terms of minimizing the variance of the importance weights, is $p(x_{t+1}|x_t, y_{t+1})$.

3.3 Auxiliary particle filtering algorithms

An alternative when the SIR algorithm performs poorly is the auxiliary particle filter (APF). The original description of the APF used the idea of auxiliary variables. The algorithm we provide motivates the APF as an importance sampling version of the exact sampling algorithm given in the previous section.

Like exact sampling, the APF consists of two steps: resampling old particles and propagating states. Unlike the exact sampling algorithm, the APF uses importance sampling when it is not possible to evaluate $p(y_{t+1}|x_t)$ or sample directly from $p(x_{t+1}|x_t, y_{t+1})$. The exact mixture weights $p(y_{t+1}|x_t)$ are approximated by an importance weight $q(y_{t+1}|x_t)$ and the posterior distribution $p(x_{t+1}|x_t, y_{t+1})$ is approximated by the importance distribution $q(x_{t+1}|x_t, y_{t+1})$. The APF algorithm is given by:

$$\text{Step 1: Compute } w\left(x_t^{(i)}\right) = \frac{q\left(y_{t+1}|x_t^{(i)}\right)}{\sum_{i=1}^{N} q\left(y_{t+1}|x_t^{(i)}\right)} \text{for } i = 1, ..., N$$

$$\text{Step 2: Draw } z\left(i\right) \sim Mult_N\left(\left\{w\left(x_t^{(i)}\right)\right\}\right) \text{ and set } x_t^{(i)} = x_t^{z(i)}$$

$$\text{Step 3: Draw } x_{t+1}^{(i)} \sim q\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right) \text{ for } i = 1, ..., N$$

$$\text{Step 4: Reweight: } \pi\left(x_{t+1}^{(i)}\right) \propto \frac{\text{target}}{\text{proposal}} = \frac{p\left(y_{t+1}|x_{t+1}^{(i)}\right) p\left(x_{t+1}^{(i)}|x_t^{(i)}\right)}{q\left(y_{t+1}|x_t^{(i)}\right) q\left(x_{t+1}^{(i)}|x_t^{(i)}, y_{t+1}\right)}.$$

The weights at the end of the algorithm are the importance sampling weights. There is no need to resampling additionally using these weights, in fact, this introduces additional Monte Carlo error.

Like exact sampling, the APF resamples first, which is important to insure that high likelihood states are propagated forward. The performance of the APF is driven by the accuracy of the importance densities. If these are poor approximations, the APF may not perform much better than the SIR algorithm, and in some extreme cases, could even perform worse. A final advantage of the APF algorithm is its flexibility, as it allows for two importance densities. This allows the algorithm to be tailored to the specific application at hand.

4 Further Reading

Research on particle filtering methods has exploded recently over the past 10 years. It is impossible to cite all of the relevant work, and we will instead focus on the initial theoretical contributions, important review papers, and applications. For textbook discussions, see the monographs by Doucet, de Freitas, and Gordon (2001) and Ristic, Arulampalam, and Gordon (2004). These books provide more details, numerous alternative algorithms and extensions to improve performance, and extensive lists of references. Cappe, Godsill, and Moulines (2007) provide a very readable and up to date review article.

The sampling/importance resampling algorithm appears in Rubin (1987) and Smith and Gelfand (1992). The foundational particle filtering algorithm appears in Gordon, Salmond, and Smith (1993). Liu and Chen (1995, 1998) provide key contributions to sequential importance sampling. Pitt and Shephard (1999) developed the auxiliary particle filter and discuss various extensions and applications. Other important contributions are in Kitigawa (1996), Hurzeler and Kunsch (1998), Carpenter, Clifford, and Fearnhead (1999), and Kunsch (2005)

For applications in economics and finance, Kim, Shephard, and Chib (1998) and Chib, Nardari and Shephard (2006) apply particle filters to univariate and multivariate stochastic volatility models. Johannes, Polson, and Stroud (2008) develop particle filters for continuous-time jump-diffusion models, with option pricing applications. Pitt (2005) discusses particle filtering approaches for maximum likelihood estimation. Fernandez-Villaverde and Rubio-Ramirez (2005) use particle filters for parameter estimation in general equilibrium macroeconomics models.

There is also a growing literature applying particle filters for sequential parameter learning and state filtering, see, for example Liu and West (2001), Storvik (2002), Fearnhead (2002), Johannes, Polson and Stroud (2005, 2006), and Johannes and Polson (2006).

References

Cappe, O., Godsill, S. and Moulines, E. (2007): An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo. *Proceedings of the IEEE* **95**, 899–924.

Carpenter, J., Clifford, P. and Fearnhead, P. (1999): An improved particle filter for nonlinear problems. *IEE Proceedings – Radar, Sonar and Navigation* **146**, 2–7.

Chib, S., Nardari, F. and Shephard, N. (2006): Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics* **134**, 341–371.

Doucet, A., de Freitas, N. and Gordon, N. (2001): *Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science*. Springer, New York.

Fearnhead, P. (2002): MCMC, sufficient statistics and particle filter. *Journal of Computuational and Graphical Statistics* **11**, 848–862.

Fernandez-Villaverde, J. and Rubio-Ramirez, J. (2005): Estimating Macroeconomic Models: A Likelihood Approach. *Working paper, University of Pennsylvania.*

Gordon, N., Salmond, D. and Smith, A.F.M. (1993): Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings* **F-140**, 107–113.

Kitagawa, G. (1996): Monte Carlo Filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics* **5**, 1–25.

Hurzeler, M. and Kunsch, H. (1998): Monte Carlo approximations for general state space models. *Journal of Computational and Graphical Statistics* **7**, 175–193.

Johannes, M. and Polson, N.G. (2006): Exact Bayesian particle filtering and parameter learning. *Working paper, University of Chicago.*

Johannes, M., Polson, N.G. and Stroud, J.R. (2008): Optimal filtering of Jump Diffusions: Extracting latent states from asset prices. *Review of Financial Studies* forthcoming.

Johannes, M., Polson, N.G. and Stroud, J.R. (2005): Sequential parameter estimation in stochastic volatility models with jumps. *Working paper, University of Chicago.*

Johannes, M., Polson, N.G. and Stroud, J.R (2006): Interacting particle systems for sequential parameter learning. *Working paper, University of Chicago.*

Kim, S., Shephard, N. and Chib, S. (1998): Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**, 361–93.

Kunsch, H. (2005): Recursive Monte Carlo filters: Algorithms and theoretical analysis. *Annals of Statistics* **33**, 1983–2021.

Liu, J. and West, M. (2001): Combined parameter and state estimation in simulation-based filtering. *In: Doucet, A. et al. (Eds.): Sequential Monte Carlo Methods in Practice* Springer, New York.

Liu, J. and Chen, R. (1995): Blind deconvolution via sequential imputations. *Journal of American Statistical Association.* **89**, 278–288.

Liu, J. and Chen, R. (1998): Sequential Monte Carlo Methods for Dynamical Systems. *Journal of the American Statistical Association* **93**, 1032–1044.

Pitt, M. (2005): Smooth particle filters for likelihood evaluation and maximisation. *Working paper, University of Warwick.*

Pitt, M. and Shephard, N. (1999): Filtering via simulation: auxiliary particle filter. *Journal of the American Statistical Association* **94**, 590–599.

Ristic, B., Arulampalam, S. and Gordon, N. (2004): *Beyond the Kalman filter: Particle filters for tracking applications.* Artech House, Boston.

Rubin, D. B. (1987): A noniterative sampling/impotance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm, comment to a paper by Tanner and Wong. *Journal of the American Statistical Association* **82**, 543–546.

Smith, A.F.M. and Gelfand, A. E. (1992): Bayesian statistics without tears. *American Statistician* **46**, 84–88.

Storvik, G. (2002): Particle filters in state space models with the presence of unknown static parameters. *IEEE Trans. on Signal Processing* **50**, 281–289.

Index