ADVANCES IN ECONOMETRICS
VOLUME 20

ECONOMETRIC ANALYSIS OF FINANCIAL AND ECONOMIC TIME SERIES - PART B

THOMAS B. FOMBY
DEK TERRELL
Editors

ECONOMETRIC ANALYSIS OF FINANCIAL AND ECONOMIC TIME SERIES

ADVANCES IN ECONOMETRICS

Series Editors: Thomas B. Fomby and R. Carter Hill

Recent Volumes:

Volume 15: Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Edited by Badi Baltagi

Volume 16: Econometric Models in Marketing, Edited by P. H. Franses and A. L. Montgomery

Volume 17: Maximum Likelihood of Misspecified Models: Twenty Years Later, Edited by Thomas B. Fomby and R. Carter Hill

Volume 18: Spatial and Spatiotemporal Econometrics, Edited by J. P. LeSage and R. Kelley Pace

Volume 19: Applications of Artificial Intelligence in Finance and Economics, Edited by J. M. Binner, G. Kendall and S. H. Chen

ADVANCES IN ECONOMETRICS VOLUME 20, PART B

ECONOMETRIC ANALYSIS OF FINANCIAL AND ECONOMIC TIME SERIES

EDITED BY

THOMAS B. FOMBY

Department of Economics, Southern Methodist University, Dallas, TX 75275

DEK TERRELL

Department of Economics, Louisiana State University, Baton Rouge, LA 70803

ELSEVIER
JAI

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER B.V.
Radarweg 29
P.O. Box 211
1000 AE Amsterdam,
The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego
CA 92101-4495
USA

ELSEVIER Ltd
The Boulevard, Langford
Lane, Kidlington
Oxford OX5 1GB
UK

ELSEVIER Ltd
84 Theobalds Road
London
WC1X 8RR
UK

© 2006 Elsevier Ltd. All rights reserved.

This work is protected under copyright by Elsevier Ltd, and the following terms and conditions apply to its use:

Photocopying
Single photocopies of single chapters may be made for personal use as allowed by national copyright laws. Permission of the Publisher and payment of a fee is required for all other photocopying, including multiple or systematic copying, copying for advertising or promotional purposes, resale, and all forms of document delivery. Special rates are available for educational institutions that wish to make photocopies for non-profit educational classroom use.

Permissions may be sought directly from Elsevier's Rights Department in Oxford, UK: phone (+44) 1865 843830, fax (+44) 1865 853333, e-mail: permissions@elsevier.com. Requests may also be completed on-line via the Elsevier homepage (http://www.elsevier.com/locate/permissions).

In the USA, users may clear permissions and make payments through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA; phone: (+1) (978) 7508400, fax: (+1) (978) 7504744, and in the UK through the Copyright Licensing Agency Rapid Clearance Service (CLARCS), 90 Tottenham Court Road, London W1P 0LP, UK; phone: (+44) 20 7631 5555; fax: (+44) 20 7631 5500. Other countries may have a local reprographic rights agency for payments.

Derivative Works
Tables of contents may be reproduced for internal circulation, but permission of the Publisher is required for external resale or distribution of such material. Permission of the Publisher is required for all other derivative works, including compilations and translations.

Electronic Storage or Usage
Permission of the Publisher is required to store or use electronically any material contained in this work, including any chapter or part of a chapter.

Except as outlined above, no part of this work may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the Publisher.

Address permissions requests to: Elsevier's Rights Department, at the fax and e-mail addresses noted above.

Notice
No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2006

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN-10: 0-7623-1273-4
ISBN-13: 978-0-7623-1273-3
ISSN: 0731-9053 (Series)

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

CONTENTS

DEDICATION *ix*

LIST OF CONTRIBUTORS *xi*

INTRODUCTION
Thomas B. Fomby and Dek Terrell *xiii*

REMARKS BY ROBERT F. ENGLE III AND SIR CLIVE W. J. GRANGER, KB
Given During Third Annual Advances in Econometrics Conference at Louisiana State University, Baton Rouge, November 5–7, 2004

GOOD IDEAS
Robert F. Engle III *xix*

THE CREATIVITY PROCESS
Sir Clive W. J. Granger, KB *xxiii*

REALIZED BETA: PERSISTENCE AND PREDICTABILITY
Torben G. Andersen, Tim Bollerslev, Francis X. Diebold and Ginger Wu *1*

ASYMMETRIC PREDICTIVE ABILITIES OF NONLINEAR MODELS FOR STOCK RETURNS: EVIDENCE FROM DENSITY FORECAST COMPARISON
Yong Bao and Tae-Hwy Lee *41*

FLEXIBLE SEASONAL TIME SERIES MODELS
Zongwu Cai and Rong Chen 63

ESTIMATION OF LONG-MEMORY TIME SERIES MODELS: A SURVEY OF DIFFERENT LIKELIHOOD-BASED METHODS
Ngai Hang Chan and Wilfredo Palma 89

BOOSTING-BASED FRAMEWORKS IN FINANCIAL MODELING: APPLICATION TO SYMBOLIC VOLATILITY FORECASTING
Valeriy V. Gavrishchaka 123

OVERLAYING TIME SCALES IN FINANCIAL VOLATILITY DATA
Eric Hillebrand 153

EVALUATING THE 'FED MODEL' OF STOCK PRICE VALUATION: AN OUT-OF-SAMPLE FORECASTING PERSPECTIVE
Dennis W. Jansen and Zijun Wang 179

STRUCTURAL CHANGE AS AN ALTERNATIVE TO LONG MEMORY IN FINANCIAL TIME SERIES
Tze Leung Lai and Haipeng Xing 205

TIME SERIES MEAN LEVEL AND STOCHASTIC VOLATILITY MODELING BY SMOOTH TRANSITION AUTOREGRESSIONS: A BAYESIAN APPROACH
Hedibert Freitas Lopes and Esther Salazar 225

ESTIMATING TAYLOR-TYPE RULES: AN UNBALANCED REGRESSION?
Pierre L. Siklos and Mark E. Wohar 239

BAYESIAN INFERENCE ON MIXTURE-OF-EXPERTS FOR ESTIMATION OF STOCHASTIC VOLATILITY
Alejandro Villagran and Gabriel Huerta *277*

A MODERN TIME SERIES ASSESSMENT OF "A STATISTICAL MODEL FOR SUNSPOT ACTIVITY" BY C. W. J. GRANGER (1957)
Gawon Yoon *297*

PERSONAL COMMENTS ON YOON'S DISCUSSION OF MY 1957 PAPER
Sir Clive W. J. Granger, KB *315*

A NEW CLASS OF TAIL-DEPENDENT TIME-SERIES MODELS AND ITS APPLICATIONS IN FINANCIAL TIME SERIES
Zhengjun Zhang *317*

This page intentionally left blank

DEDICATION

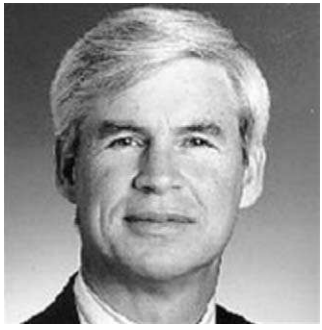
Volume 20 of *Advances in Econometrics* is dedicated to Rob Engle and Sir Clive Granger, winners of the 2003 Nobel Prize in Economics, for their many valuable contributions to the econometrics profession. The Royal Swedish Academy of Sciences cited Rob "for methods of analyzing economic time series with time-varying volatility (ARCH)" while Clive was cited "for methods of analyzing economic time series with common trends (cointegration)." Of course, these citations are meant for public consumption but we specialists in time series analysis know their contributions go far beyond these brief citations. Consider *some* of Rob's other contributions to our literature: Aggregation of Time Series, Band Spectrum Regression, Dynamic Factor Models, Exogeneity, Forecasting in the Presence of Cointegration, Seasonal Cointegration, Common Features, ARCH-M, Multivariate GARCH, Analysis of High Frequency Data, and CAViaR. *Some* of Sir Clive's additional contributions include Spectral Analysis of Economic Time Series, Bilinear Time Series Models, Combination Forecasting, Spurious Regression, Forecasting Transformed Time Series, Causality, Aggregation of Time Series, Long Memory, Extreme Bounds, Multi-Cointegration, and Non-linear Cointegration. No doubt, their Nobel Prizes are richly deserved. And the 48 authors of the two parts of this volume think likewise. They have authored some very fine papers that contribute nicely to the same literature that Rob's and Clive's research helped build.

For more information on Rob's and Clive's Nobel prizes you can go to the Nobel Prize website http://nobelprize.org/economics/laureates/2003/index.html. In addition to the papers that are contributed here, we are publishing remarks by Rob and Clive on the nature of innovation in econometric research that were given during the Third Annual *Advances in Econometrics* Conference at Louisiana State University in Baton Rouge, November 5–7, 2004. We think you will enjoy reading their remarks. You come away with the distinct impression that, although they may claim they were "lucky" or "things just happened to fall in place," having the orientation of building models that solve practical problems has been an orientation that served them and our profession very well.

We hope the readers of this two-part volume enjoy its contents. We feel fortunate to have had the opportunity of working with these fine authors and putting this volume together.

Thomas B. Fomby
Department of Economics
Southern Methodist University
Dallas, Texas 75275

Dek Terrell
Department of Economics
Louisiana State University
Baton Rouge, Louisiana 70803



Robert F. Engle III
2003 Nobel Prize Winner
in Economics

Sir Clive W. J. Granger, Knight's designation (KB)
2003 Nobel Prize Winner
in Economics

LIST OF CONTRIBUTORS

Torben G. Andersen	Department of Finance, Kellogg School of Management, Northwestern University, IL, USA
Yong Bao	Department of Economics, University of Texas, TX, USA
Tim Bollerslev	Department of Economics, Duke University, NC, USA
Zongwu Cai	Department of Mathematics & Statistics, University of North Carolina, NC, USA
Ngai Hang Chan	Department of Statistics, The Chinese University of Hong Kong, New Territories, Hong Kong
Rong Chen	Department of Information and Decision Sciences, College of Business Administration, The University of Illinois at Chicago, IL, USA
Francis X. Diebold	Department of Economics, University of Pennsylvania, PA, USA
Robert F. Engle III	Stern School of Business, New York University, NY, USA
Valeriy V. Gavrishchaka	Alexandra Investment Management, LLC, NY, USA
Sir Clive W. J. Granger, KB	Department of Economics, University of California at San Diego, CA, USA
Eric Hillebrand	Department of Economics, Louisiana State University, LA, USA
Gabriel Huerta	Department of Mathematics and Statistics, University of New Mexico, NM, USA

Dennis W. Jansen	Department of Economics, Texas A&M University, TX, USA
Tze Leung Lai	Department of Statistics, Stanford University, CA, USA
Tae-Hwy Lee	Department of Economics, University of California, CA, USA
Hedibert Freitas Lopes	Graduate School of Business, University of Chicago, IL, USA
Wilfredo Palma	Departamento de Estadistica, Pontificia Universidad Catolica De Chile (PUC), Santiago, Chile
Esther Salazar	Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
Pierre L. Siklos	Department of Economics, Wilfrid Laurier University, Ontario, Canada
Alejandro Villagran	Department of Mathematics and Statistics, University of New Mexico, NM, USA
Zijun Wang	Private Enterprise Research Center, Texas A&M University, TX, USA
Mark E. Wohar	Department of Economics, University of Nebraska–Omaha, NE, USA
Ginger Wu	Department of Banking and Finance, University of Georgia, GA, USA
Haipeng Xing	Department of Statistics, Columbia University, NY, USA
Gawon Yoon	School of Economics, Kookmin University, Seoul, South Korea
Zhengjun Zhang	Department of Statistics, University of Wisconsin-Madison, WI, USA

INTRODUCTION

Thomas B. Fomby and Dek Terrell

The editors are pleased to offer the following papers to the reader in recognition and appreciation of the contributions to our literature made by Robert Engle and Sir Clive Granger, winners of the 2003 Nobel Prize in Economics. Please see the previous dedication page of this volume. The basic themes of this part of Volume 20 of *Advances in Econometrics* are time-varying betas of the capital asset pricing model, analysis of predictive densities of nonlinear models of stock returns, modeling multivariate dynamic correlations, flexible seasonal time series models, estimation of long-memory time series models, the application of the technique of boosting in volatility forecasting, the use of different time scales in Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH) modeling, out-of-sample evaluation of the 'Fed Model' in stock price valuation, structural change as an alternative to long memory, the use of smooth transition autoregressions in stochastic volatility modeling, the analysis of the "balancedness" of regressions analyzing Taylor-type rules of the Fed Funds rate, a mixture-of-experts approach for the estimation of stochastic volatility, a modern assessment of Clive's first published paper on sunspot activity, and a new class of models of tail-dependence in time series subject to jumps. Of course, we are also pleased to include Rob's and Clive's remarks on their careers and their views on innovation in econometric theory and practice that were given at the Third Annual *Advances in Econometrics* Conference held at Louisiana State University, Baton Rouge, on November 5–7, 2004.

Let us briefly review the specifics of the papers presented here. In the first paper, "Realized Beta: Persistence and Predictability," Torben Andersen, Tim Bollerslev, Francis Diebold, and Jin Wu review the literature on the one-factor Capital Asset Pricing Model (CAPM) for the purpose of coming to a better understanding of the variability of the betas of such models. They do this by flexibly modeling betas as the ratio of the integrated stock and market return covariance and integrated market variance in a way that allows, but does not impose, fractional integration and/or cointegration.

They find that, although the realized variances and covariances fluctuate widely and are highly persistent and predictable, the realized betas, which are simple nonlinear functions of the realized variances and covariances, display much less persistence and predictability. They conclude that the constant beta CAPM, as bad as it may be, is nevertheless not as bad as some popular conditional CAPMs. Their paper provides some very useful insight into why allowing for time-varying betas may do more harm than good when estimated from daily data. They close by sketching an interesting framework for future research using high-frequency intraday data to improve the modeling of time-varying betas.

Yong Bao and Tae-Hwy Lee in their paper "Asymmetric Predictive Abilities of Nonlinear Models for Stock Returns: Evidence from Density Forecast Comparison" investigate the nonlinear predictability of stock returns when the density forecasts are evaluated and compared instead of the conditional mean point forecasts. They use the Kullback–Leibler Information Criterion (KLIC) divergence measure to characterize the extent of misspecification of a forecast model. Their empirical findings suggest that the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric in the sense that the right tails of the return series are predictable via many of the nonlinear models while they find no such evidence for the left tails or the entire distribution.

Zongwu Cai and Rong Chen introduce a new class of flexible seasonal time series models to characterize trend and seasonal variation in their paper "Flexible Seasonal Time Series Models." Their model consists of a common trend function over periods and additive individual trend seasonal functions that are specific to each season within periods. A local linear approach is developed to estimate the common trend and seasonal trend functions. The consistency and asymptotic normality of the proposed estimators are established under weak α-mixing conditions and without specifying the error distribution. The proposed methodologies are illustrated with a simulated example and two economic and financial time series, which exhibit nonlinear and nonstationary behavior.

In "Estimation of Long-Memory Time Series Models: A Survey of Different Likelihood-Based Methods" Ngai Hang Chan and Wilfredo Palma survey the various likelihood-based techniques for analyzing long memory in time series data. The authors classify these methods into the following categories: Exact maximum likelihood methods, maximum likelihood methods based on autoregressive approximations, Whittle estimates, Whittle estimates with autoregressive truncation, approximate estimates based on the Durbin–Levinson algorithm, state–space based estimates for

the autoregressive fractionally integrated moving average (ARFIMA) models, and estimation of stochastic volatility models. Their review provides a succinct survey of these methodologies as well as an overview of the important related problems such as the maximum likelihood estimation with missing data, influence of subsets of observations on estimates, and the estimation of seasonal long-memory models. Performances and asymptotic properties of these techniques are compared and examined. Interconnections and finite sample performances among these procedures are studied and applications to financial time series of these methodologies are discussed.

In ''Boosting-Based Frameworks in Financial Modeling: Application to Symbolic Volatility Forecasting'' Valeriy Gavrishchaka suggests that Boosting (a novel ensemble learning technique) can serve as a simple and robust framework for combining the best features of both analytical and data-driven models and more specifically discusses how Boosting can be applied for typical financial and econometric applications. Furthermore, he demonstrates some of the capabilities of Boosting by showing how a Boosted collection of GARCH-type models for the IBM stock time series can be used to produce more accurate forecasts of volatility than both the best single model of the collection and the widely used GARCH(1,1) model.

Eric Hillebrand studies different generalizations of GARCH that allow for several time scales in his paper ''Overlaying Time Scales in Financial Volatility Data.'' In particular he examines the nature of the volatility in four measures of the U.S. stock market (S&P 500, Dow Jones Industrial Average, CRSP equally weighted index, and CRSP value-weighted index) as well as the exchange rate of the Japanese Yen against the U.S. Dollar and the U.S. federal funds rate. In addition to analyzing these series using the conventional ARCH and GARCH models he uses three models of multiple time scales, namely, fractional integration, two-scale GARCH, and wavelet analysis in conjunction with Heterogeneous ARCH (HARCH). Hillebrand finds that the conventional ARCH and GARCH models miss the important short correlation time scale in the six series. Based on a holding sample test, the multiple time scale models, although offering an improvement over the conventional ARCH and GARCH models, still did not completely model the short correlation structure of the six series. However, research in extending volatility models in this way appears to be promising.

The ''Fed Model'' postulates a cointegrating relationship between the equity yield on the S&P 500 and the bond yield. In their paper, ''Evaluating the 'Fed Model' of Stock Price Valuation: An Out-of-Sample Forecasting Perspective,'' Dennis Jansen and Zijun Wang evaluate the Fed Model as a

vector-error correction forecasting model for stock prices and for bond yields. They compare out-of-sample forecasts of each of these two variables from a univariate model and various versions of the Fed Model including both linear and nonlinear vector error correction models. They find that for stock prices the Fed Model improves on the univariate model for longer-horizon forecasts, and the nonlinear vector-error correction model performs even better than its linear version.

In their paper, "Structural Change as an Alternative to Long Memory in Financial Time Series," Tze Leung Lai and Haipeng Xing note that volatility persistence in GARCH models and spurious long memory in autoregressive models may arise if the possibility of structural changes is not incorporated in the time series model. Therefore, they propose a structural change model that allows changes in the volatility and regression parameters at unknown times and with unknown changes in magnitudes. Their model is a hidden Markov model in which the volatility and regression parameters can continuously change and are estimated by recursive filters. As their hidden Markov model involves gamma-normal conjugate priors, there are explicit recursive formulas for the optimal filters and smoothers. Using NASDAQ weekly return data, they show how the optimal structural change model can be applied to segment financial time series by making use of the estimated probabilities of structural breaks.

In their paper, "Time Series Mean Level and Stochastic Volatility Modeling by Smooth Transition Autoregressions: A Bayesian Approach," Hedibert Lopes and Esther Salazar propose a Bayesian approach to model the level and variance of financial time series based on a special class of nonlinear time series models known as the logistic smooth transition autoregressive (LSTAR) model. They propose a Markov Chain Monte Carlo (MCMC) algorithm for the levels of the time series and then adapt it to model the stochastic volatilities. The LSTAR order of their model is selected by the three information criteria Akaike information criterion (AIC), Bayesian information criterion (BIC), and Deviance information criteria (DIC). They apply their algorithm to one synthetic and two real-time series, namely the Canadian Lynx data and the SP500 return series, and find the results encouraging when modeling both the levels and the variance of univariate time series with LSTAR structures.

Relying on Robert Engle's and Clive Granger's many and varied contributions to econometrics analysis, Pierre Siklos and Mark Wohar examine some key econometric considerations involved in estimating Taylor-type rules for U.S. data in their paper "Estimating Taylor-Type Rules: An Unbalanced Regression?" They focus on the roles of unit roots, cointegration,

structural breaks, and nonlinearities to make the case that most existing estimates are based on unbalanced regressions. A variety of their estimates reveal that neglecting the presence of cointegration results in the omission of a necessary error correction term and that Fed reactions during the Greenspan era appear to have been asymmetric. They further argue that error correction and nonlinearities may be one way to estimate Taylor rules over long samples when the underlying policy regime may have changed significantly.

Alejandro Villagran and Gabriel Huerta propose a Bayesian Mixture-of-Experts (ME) approach to estimating stochastic volatility in time series in their paper "Bayesian Inference on Mixture-of-Experts for Estimation of Stochastic Volatility." They use as their "experts" the ARCH, GARCH, and EGARCH models to analyze the stochastic volatility in the U.S. dollar/German Mark exchange rate and conduct a study of the volatility of the Mexican stock market (IPC) index using the Dow Jones Industrial (DJI) index as a covariate. They also describe the estimation of predictive volatilities and their corresponding measure of uncertainty given by a Bayesian credible interval using the ME approach. In the applications they present, it is interesting to see how the posterior probabilities of the "experts" change over time and to conjecture why the posterior probabilities changed as they did.

Sir Clive Granger published his first paper "A Statistical Model for Sunspot Activity" in 1957 in the prestigious *Astrophysical Journal*. As a means of recognizing Clive's many contributions to the econometric analysis of time series and celebrating the near 50th anniversary of his first publication, one of his students and now professor, Gawon Yoon, has written a paper that provides a modern time series assessment of Clive's first paper. In "A Modern Time Series Assessment of 'A Statistical Model for Sunspot Activity' by C. W. J. Granger (1957)," Yoon reviews Granger's statistical model of sunspots containing two parameters representing an amplitude factor and the occurrence of minima, respectively. At the time Granger's model accounted for about 85% of the total variation in the sunspot data. Interestingly, Yoon finds that, in the majority, Granger's model quite nicely explains the more recent occurrence of sunspots despite the passage of time. Even though it appears that some of the earlier observations that Granger had available were measured differently from later sunspot numbers, Granger's simple two-parameter model still accounts for more than 80% of the total variation in the extended sunspot data. This all goes to show (as Sir Clive would attest) that simple models can also be useful models.

With respect to Yoon's review of Granger's paper, Sir Clive was kind enough to offer remarks that the editors have chosen to publish immediately

following Yoon's paper. In reading Clive's delightful remarks we come to know (or some of us remember, depending on your age) how difficult it was to conduct empirical analysis in the 1950s and 1960s. As Clive notes, "Trying to plot by hand nearly two hundred years of monthly data is a lengthy task!" So econometric researchers post-1980 have many things to be thankful for, not the least of which is fast and inexpensive computing. Nevertheless, Clive and many other young statistical researchers at the time were undaunted. They were convinced that quantitative research was important and a worthwhile endeavor regardless of the expense of time and they set out to investigate what the naked eye could not detect. You will enjoy reading Clive's remarks knowing that he is always appreciative of the comments and suggestions of colleagues and that he is an avid supporter of best practices in statistics and econometrics.

In the final paper of Part B of Volume 20, "A New Class of Tail-Dependent Time Series Models and Its Applications in Financial Time Series," Zhengjun Zhang proposes a new class of models to determine the order of lag-k tail dependence in financial time series that exhibit jumps. His base model is a specific class of maxima of moving maxima processes (M3 processes). Zhang then improves on his base model by allowing for possible asymmetry between positive and negative returns. His approach adopts a hierarchical model structure. First you apply, say GARCH(1,1), to get estimated standard deviations, then based on standardized returns, you apply M3 and Markov process modeling to characterize the tail dependence in the time series. Zhang demonstrates his model and his approach using the S&P 500 Index. As he points out, estimates of the parameters of the proposed model can be used to compute the value at risk (VaR) of the investments whose returns are subject to jump processes.

GOOD IDEAS

Robert F. Engle III

The Nobel Prize is given for good ideas – very good ideas. These ideas often shape the direction of research for an academic discipline. These ideas are often accompanied by a great deal of work by many researchers.

Most good ideas don't get prizes but they are the centerpieces of our research and our conferences. At this interesting *Advances in Econometrics* conference hosted by LSU, we've seen lots of new ideas, and in our careers we have all had many good ideas. I would like to explore where they come from and what they look like.

When I was growing up in suburban Philadelphia, my mother would sometimes take me over to Swarthmore College to the Physics library. It was a small dusty room with windows out over a big lawn with trees. The books cracked when I opened them; they smelled old and had faded gold letters on the spine. This little room was exhilarating. I opened books by the famous names in physics and read about quantum mechanics, elementary particles and the history of the universe. I didn't understand too much but kept piecing together my limited ideas. I kept wondering whether I would understand these things when I was older and had studied in college or graduate school. I developed a love of science and the scientific method. I think this is why I studied econometrics; it is the place where theory meets reality. It is the place where data on the economy tests the validity of economic theory.

Fundamentally I think good ideas are simple. In Economics, most ideas can be simplified until they can be explained to non-specialists in plain language. The process of simplifying ideas and explaining them is extremely important. Often the power of the idea comes from simplification of a collection of complicated and conflicting research. The process of distilling out the simple novel ingredient is not easy at all and often takes lots of fresh starts and numerical examples. Discouragingly, good ideas boiled down to their essence may seem trivial. I think this is true of ARCH and Cointegration and many other Nobel citations. But, I think we should not be

offended by this simplicity, but rather we should embrace it. Of course it is easy to do this after 20 years have gone by; but the trick is to recognize good ideas early. Look for them at seminars or when reading or refereeing or editing.

Good ideas generalize. A good idea, when applied to a new situation, often gives interesting insights. In fact, the implications of a good idea may be initially surprising. Upon reflection, the implications may be of growing importance. If ideas translated into other fields give novel interpretations to existing problems, this is a measure of their power.

Often good ideas come from examining one problem from the point of view of another. In fact, the ARCH model came from such an analysis. It was a marriage of theory, time series and empirical evidence. The role of uncertainty in rational expectations macroeconomics was not well developed, yet there were theoretical reasons why changing uncertainty could have real effects. From a time series point of view a natural solution to modeling uncertainty was to build conditional models of variance rather than the more familiar unconditional models. I knew that Clive's test for bilinearity based on the autocorrelation of squared residuals was often significant in macroeconomic data, although I suspected that the test was also sensitive to other effects such as changing variances. The idea for the ARCH model came from combining these three observations to get an autoregressive model of conditional heteroskedasticity.

Sometimes a good idea can come from attempts to disprove proposals of others. Clive traces the origin of cointegration to his attempt to disprove a David Hendry conjecture that a linear combination of the two integrated series could be stationary. From trying to show that this was impossible, Clive proved the Granger Representation theorem that provides the fundamental rationale for error correction models in cointegrated systems.

My first meeting in Economics was the 1970 World Congress of the Econometric Society in Cambridge England. I heard many of the famous economists of that generation explain their ideas. I certainly did not understand everything but I wanted to learn it all. I gave a paper at this meeting at a session organized by Clive that included Chris Sims and Phoebus Dhrymes. What a thrill. I have enjoyed European meetings of the Econometric Society ever since.

My first job was at MIT. I had a lot of chances to see good ideas; particularly good ideas in finance. Myron Scholes and Fischer Black were working on options theory and Bob Merton was developing continuous time finance. I joined Franco Modigliani and Myron on Michael Brennan's dissertation committee where he was testing the CAPM. Somehow I missed

the opportunity to capitalize on these powerful ideas and it was only many years later that I moved my research in this direction.

I moved to UCSD in 1976 to join Clive Granger. We studied many fascinating time series problems. Mark Watson was my first PhD student at UCSD. The ARCH model was developed on sabbatical at LSE, and when I returned, a group of graduate students contributed greatly to the development of this research. Tim Bollerslev and Dennis Kraft were among the first, Russ Robins and Jeff Wooldridge and my colleague David Lilien were instrumental in helping me think about the finance applications. The next 20 years at UCSD were fantastic in retrospect. I don't think we knew at the time how we were moving the frontiers in econometrics. We had great visitors and faculty and students and every day there were new ideas.

These ideas came from casual conversations and a relaxed mind. They came from brainstorming on the blackboard with a student who was looking for a dissertation topic. They came from "Econometrics Lunch" when we weren't talking about gossip in the profession. Universities are incubators of good ideas. Our students come with good ideas, but they have to be shaped and interpreted. Our faculties have good ideas, which they publish and lecture on around the world. Our departments and universities thrive on good ideas that make them famous places for study and innovation. They also contribute to spin-offs in the private sector and consulting projects. Good ideas make the whole system work and it is so important to recognize them in all their various forms and reward them.

As a profession we are very protective of our ideas. Often the origin of the idea is disputable. New ideas may have only a part of the story that eventually develops; who gets the credit? While such disputes are natural, it is often better in my opinion to recognize previous contributions and stand on their shoulders thereby making your own ideas even more important. I give similar advice for academics who are changing specialties; stand with one foot in the old discipline and one in the new. Look for research that takes your successful ideas from one field into an important place in a new discipline.

Here are three quotations that I think succinctly reflect these thoughts.

- *"The universe is full of magical things, patiently waiting for our wits to grow sharper"*
  Eden Philpotts
- *"To see what is in front of one's nose requires a constant struggle."*
  George Orwell
- *"To select well among old things is almost equal to inventing new ones"*
  Nicolas Charles Trublet

There is nothing in our chosen career that is as exhilarating as having a good idea. But a very close second is seeing someone develop a wonderful new application from your idea. The award of the Nobel Prize to Clive and me for our work in time series is an honor to all of the authors who contributed to the conference and to this volume. I think the prize is really given to a field and we all received it. This gives me so much joy. And I hope that someone in this volume will move forward to open more doors with powerful new ideas, and receive her own Nobel Prize.

Robert F. Engle III
Remarks Given at Third Annual Advances in Econometrics Conference
Louisiana State University
Baton Rouge, Louisiana
November 5–7, 2004

THE CREATIVITY PROCESS

Sir Clive W. J. Granger, KB

In 1956, I was searching for a Ph.D. topic and I selected time series analysis as being an area that was not very developed and was potentially interesting. I have never regretted that choice. Occasionally, I have tried to develop other interests but after a couple of years away I would always return to time series topics where I am more comfortable.

I have never had a long-term research topic. What I try to do is to develop new ideas, topics, and models, do some initial development, and leave the really hard, rigorous stuff to other people. Some new topics catch on quickly and develop a lot of citations (such as cointegration), others are initially ignored but eventually become much discussed and applied (causality, as I call it), some develop interest slowly but eventually deeply (fractionally integrated processes), some have long term, steady life (combination of forecasts), whereas others generate interest but eventually vanish (bilinear models, spectral analysis).

The ideas come from many sources, by reading literature in other fields, from discussions with other workers, from attending conferences (time distance measure for forecasts), and from general reading. I will often attempt to take a known model and generalize and expand it in various ways. Quite frequently these generalizations turn out not to be interesting; I have several examples of general I(d) processes where d is not real or not finite. The models that do survive may be technically interesting but they may not prove useful with economic data, providing an example of a so-called "empty box," bilinear models, and I(d), d non-integer could be examples.

In developing these models one is playing a game. One can never claim that a new model will be relevant, only that it might be. Of course, when using the model to generate forecasts, one has to assume that the model is correct, but one must not forget this assumption. If the model is correct, the data will have certain properties that can be proved, but it should always be remembered that other models may generate the same properties,

for example I(d), d a fraction, and break processes can give similar "long memory" autocorrelations. Finding properties of data and then suggesting that a particular model will have generated the data is a dangerous game.

Of course, once the research has been done one faces the problem of publication. The refereeing process is always a hassle. I am not convinced that delaying an interesting paper (I am not thinking of any of my own here) by a year or more to fix a few minor difficulties is actually helping the development of our field. Rob and I had initial rejections of some of our best joint papers, including the one on cointegration. My paper on the typical spectral shape took over three and a half years between submission and publication, and it is a very short paper.

My favorite editors' comment was that "my paper was not very good (correct) but is was very short," and as they just had that space to fill they would accept. My least favorite comment was a rejection of a paper with Paul Newbold because "it has all been done before." As we were surprised at this we politely asked for citations. The referee had no citations, he just thought that must have been done before. The paper was published elsewhere.

For most of its history time series theory considered conditional means, but later conditional variances. The next natural development would be conditional quantiles, but this area is receiving less attention than I expected. The last stages are initially conditional marginal distributions, and finally conditional multivariate distributions. Some interesting theory is starting in these areas but there is an enormous amount to be done.

The practical aspects of time series analysis are rapidly changing with improvements in computer performance. Now many, fairly long series can be analyzed jointly. For example, Stock and Watson (1999) consider over 200 macro series. However, the dependent series are usually considered individually, whereas what we are really dealing with is a sample from a 200-dimensional multivariate distribution, assuming the processes are jointly stationary. How to even describe the essential features of such a distribution, which is almost certainly non-Gaussian, in a way that is useful to economists and decision makers is a substantial problem in itself.

My younger colleagues sometimes complain that we old guys solved all the interesting easy questions. I do not think that was ever true and is not true now. The higher we stand the wider our perspective; I hope that Rob and I have provided, with many others, a suitable starting point for the future study in this area.

REFERENCE

Stock, J. H., & Watson, M. W. (1999). A Comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: R. F. Engle & H. White (Eds), *Cointegration, causality, and forecasting: A festschrift in honour of Clive W. J. Granger*. Oxford: Oxford University Press.

Sir Clive W. J. Granger, KB
Remarks Read at Third Annual Advances in Econometrics Conference
Louisiana State University
Baton Rouge, Louisiana
November 5–7, 2004

This page intentionally left blank

REALIZED BETA: PERSISTENCE AND PREDICTABILITY ☆

Torben G. Andersen, Tim Bollerslev,
Francis X. Diebold and Ginger Wu

ABSTRACT

A large literature over several decades reveals both extensive concern with the question of time-varying betas and an emerging consensus that betas are in fact time-varying, leading to the prominence of the conditional CAPM. Set against that background, we assess the dynamics in realized betas, vis-à-vis the dynamics in the underlying realized market variance and individual equity covariances with the market. Working in the recently popularized framework of realized volatility, we are led to a framework of nonlinear fractional cointegration: although realized variances and covariances are very highly persistent and well approximated as fractionally integrated, realized betas, which are simple nonlinear functions of those realized variances and covariances, are less persistent and arguably best modeled as stationary I(0) processes. We conclude by drawing implications for asset pricing and portfolio management.

☆We dedicate this paper to Clive W. J. Granger, a giant of modern econometrics, on whose broad shoulders we are fortunate to stand. This work was supported by the National Science Foundation and the Guggenheim Foundation.

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 1–39
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20020-8

1. INTRODUCTION

One of the key insights of asset pricing theory is also one of the simplest: only systematic risk should be priced. Perhaps not surprisingly, however, there is disagreement as to the sources of systematic risk. In the one-factor capital asset pricing model (CAPM), for example, systematic risk is determined by covariance with the market (Sharpe, 1963; Lintner, 1965a, b), whereas, in more elaborate pricing models, additional empirical characteristics such as firm size and book-to-market are seen as proxies for another set of systematic risk factors (Fama & French, 1993).[1]

As with most important scientific models, the CAPM has been subject to substantial criticism (e.g., Fama & French, 1992). Nevertheless, to paraphrase Mark Twain, the reports of its death are greatly exaggerated. In fact, the one-factor CAPM remains alive and well at the frontier of both academic research and industry applications, for at least two reasons. First, recent work reveals that it often works well – despite its wrinkles and warts – whether in traditional incarnations (e.g., Ang & Chen, 2003) or more novel variants (e.g., Cohen, Polk, & Vuolteenaho, 2002; Campbell & Vuolteenaho, 2004). Second, competing multi-factor pricing models, although providing improved statistical fit, involve factors whose economic interpretations in terms of systematic risks remain unclear, and moreover, the stability of empirically motivated multi-factor asset pricing relationships often appears tenuous when explored with true out-of-sample data, suggesting an element of data mining.[2]

In this paper, then, we study the one-factor CAPM, which remains central to financial economics nearly a half century after its introduction. A key question within this setting is whether stocks' systematic risks, as assessed by their correlations with the market, are constant over time – i.e., whether stocks' market betas are constant. And if betas are not constant, a central issue becomes how to understand and formally characterize their persistence and predictability vis-à-vis their underlying components.

The evolution of a large literature over several decades reveals both extensive concern with this question and, we contend, an eventual implicit consensus that betas *are* likely time-varying.[3] Several pieces of evidence support our contention. First, leading texts echo it. For example, Huang and Litzenberger (1988) assert that "It is unlikely that risk premiums and betas on individual assets are stationary over time" (p. 303). Second, explicitly dynamic betas are often modeled nonstructurally via time-varying parameter regression, in a literature tracing at least to the early "return to normality" model of Rosenberg (1973), as implemented in the CAPM by Schaefer Broaley, Hodges, and Thomas (1975). Third, even in the absence of

explicit allowance for time-varying betas, the CAPM is typically estimated using moving estimation windows, usually of 5–10 years, presumably to guard against beta variation (e.g., Fama, 1976; Campbell, Lo, & MacKinlay, 1997). Fourth, theoretical and empirical inquiries in asset pricing are often undertaken in conditional, as opposed to unconditional, frameworks, the essence of which is to allow for time-varying betas, presumably because doing so is viewed as necessary for realism.

The motivation for the conditional CAPM comes from at least two sources. First, from a theoretical perspective, financial economic considerations suggest that betas may vary with conditioning variables, an idea developed theoretically and empirically in a large literature that includes, among many others, Dybvig and Ross (1985), Hansen and Richard (1987), Ferson, Kandel, and Stambaugh (1987), Ferson and Harvey (1991), Jagannathan and Wang (1996), and Wang (2003).[4] Second, from a different and empirical perspective, the financial econometric volatility literature (see Andersen, Bollerslev, & Diebold, 2005, for a recent survey) has provided extensive evidence of wide fluctuations and high persistence in asset market conditional variances, and in individual equity conditional covariances with the market. Thus, even from a purely statistical viewpoint, market betas, which are ratios of time-varying conditional covariances and variances, might be expected to display persistent fluctuations, as in Bollerslev, Engle, and Wooldridge (1988). In fact, unless some special cancellation occurs – in a way that we formalize – betas would inherit the persistence features that are so vividly present in their constituent components.

Set against this background, we assess the dynamics in betas vis-à-vis the widely documented persistent dynamics in the underlying variance and covariances. We proceed as follows: In Section 2, we sketch the framework, both economic and econometric, in which our analysis is couched. In Section 3, we present the empirical results with an emphasis on analysis of persistence and predictability. In Section 4, we formally assess the uncertainty in our beta estimates. In Section 5, we offer summary, conclusions, and directions for future research.

2. THEORETICAL FRAMEWORK

Our approach has two key components. First, in keeping with the recent move toward nonparametric volatility measurement, we cast our analysis within the framework of realized variances and covariances, or equivalently, empirical quadratic variation and covariation. That is, we do not entertain a null

hypothesis of period-by-period constant betas, but instead explicitly allow for continuous evolution in betas. Our "realized betas" are (continuous-record) consistent for realizations of the underlying ratio between the integrated stock and market return covariance and the integrated market variance.[5] Second, we work in a flexible econometric framework that allows for – without imposing – fractional integration and/or cointegration between the market variance and individual equity covariances with the market.

2.1. Realized Quarterly Variances, Covariances, and Betas

We provide estimates of quarterly betas, based on nonparametric realized quarterly market variances and individual equity covariances with the market. The quarterly frequency is appealing from a substantive financial economic perspective, and it also provides a reasonable balance between efficiency and robustness to microstructure noise. Specifically, we produce our quarterly estimates using underlying daily returns, as in Schwert (1989), so that the sampling frequency is quite high relative to the quarterly horizon of interest, yet low enough so that contamination by microstructure noise is not a serious concern for the highly liquid stocks that we study. The daily frequency further allows us to utilize a long sample of data, which is not available when sampling more frequently.

Suppose that the logarithmic $N \times 1$ vector price process, p_t, follows a multivariate continuous-time stochastic volatility diffusion,

$$\mathrm{d}p_t = \mu_t \mathrm{d}t + \Omega_t \mathrm{d}W_t \tag{1}$$

where W_t denotes a standard N-dimensional Brownian motion, and both the process for the $N \times N$ positive definite diffusion matrix, Ω_t, and the N-dimensional instantaneous drift, μ_t, are strictly stationary and jointly independent of the W_t process. For our purposes it is helpful to think of the Nth element of p_t as containing the log price of the market and the ith element of p_t as containing the log price of the ith individual stock included in the analysis, so that the corresponding covariance matrix contains both the market variance, say $\sigma^2_{M,t} = \Omega_{(NN),t}$, and the individual equity covariance with the market, $\sigma_{iM,t} = \Omega_{(iN),t}$. Then, conditional on the sample path realization of μ_t and Ω_t, the distribution of the continuously compounded h-period return, $r_{t+h,h} \equiv p_{t+h} - p_t$, is

$$r_{t+h,h} | \sigma\left\{\mu_{t+\tau}, \Omega_{t+\tau}\right\}_{\tau=0}^{h} \sim N\left(\int_0^h \mu_{t+\tau}\mathrm{d}\tau,\ \int_0^h \Omega_{t+\tau}\mathrm{d}\tau\right) \tag{2}$$

where $\sigma\{\mu_{t+\tau}, \Omega_{t+\tau}\}_{\tau=0}^{h}$ denotes the σ-field generated by the sample paths of $\mu_{t+\tau}$ and $\Omega_{t+\tau}$, for $0 \leq \tau \leq h$. The integrated diffusion matrix $\int_0^h \Omega_{t+\tau} d\tau$, therefore provides a natural measure of the true latent h-period volatility.[6] The requirement that the innovation process, W_t, is independent of the drift and diffusion processes is rather strict and precludes, for example, the asymmetric relations between return innovations and volatility captured by the so-called leverage or volatility feedback effects. However, from the results in Meddahi (2002), Barndorff-Nielsen and Shephard (2003), and Andersen, Bollerslev, and Meddahi (2004), we know that the continuous-record asymptotic distribution theory for the realized covariation continues to provide an excellent approximation for empirical high-frequency realized volatility measures.[7] As such, even if the conditional return distribution result (2) does not apply in full generality, the evidence presented below, based exclusively on the realized volatility measures, remains trustworthy in the presence of asymmetries in the return innovation–volatility relations.

By the theory of quadratic variation, we have that under weak regularity conditions, and regardless of the presence of leverage or volatility feedback effects, that

$$\sum_{j=1,\ldots,\lfloor h/\Delta \rfloor} r_{t+j\cdot\Delta,\Delta} \cdot r'_{t+j\cdot\Delta,\Delta} - \int_0^h \Omega_{t+\tau} d\tau \to 0 \tag{3}$$

almost surely for all t as the sampling frequency of the returns increases, or $\Delta \to 0$. Thus, by summing sufficiently finely sampled high-frequency returns, it is possible to construct ex-post *realized* volatility measures for the integrated latent volatilities that are asymptotically free of measurement error. This contrasts sharply with the common use of the cross-product of the h-period returns, $r_{t+h,h} \cdot r'_{t+h,h}$, as a simple ex post (co)variability measure. Although the squared return (innovation) over the forecast horizon provides an unbiased estimate for the integrated volatility, it is an extremely noisy estimator, and predictable variation in the true latent volatility process is typically dwarfed by measurement error. Moreover, for longer horizons any conditional mean dependence will tend to contaminate this variance measure. In contrast, as the sampling frequency is lowered, the impact of the drift term vanishes, thus effectively annihilating the mean.

These assertions remain valid if the underlying continuous time process in Eq. (1) contains jumps, so long as the price process is a special semimartingale, which will hold if it is arbitrage-free. Of course, in this case the limit of the summation of the high-frequency returns will involve an additional

jump component, but the interpretation of the sum as the realized h-period return volatility remains intact.

Finally, with the realized market variance and realized covariance between the market and the individual stocks in hand, we can readily define and empirically construct the individual equity "realized betas." Toward that end, we introduce some formal notation. Using an initial subscript to indicate the corresponding element of a vector, we denote the realized market volatility by

$$\hat{v}^2_{M,t,t+h} = \sum_{j=1,\ldots,[h/\Delta]} \boldsymbol{r}^2_{(N),t+j\cdot\Delta,\Delta} \tag{4}$$

and we denote the realized covariance between the market and the ith individual stock return by

$$\hat{v}_{iM,t,t+h} = \sum_{j=1,\ldots,[h/\Delta]} \boldsymbol{r}_{(i),t+j\cdot\Delta,\Delta} \cdot \boldsymbol{r}_{(N),t+j\cdot\Delta,\Delta} \tag{5}$$

We then define the associated realized beta as

$$\hat{\beta}_{i,t,t+h} = \frac{\hat{v}_{iM,t,t+h}}{\hat{v}^2_{M,t,t+h}} \tag{6}$$

Under the assumptions invoked for Eq. (1), this realized beta measure is consistent for the true underlying integrated beta in the following sense:

$$\hat{\beta}_{i,t,t+h} \to \beta_{i,t,t+h} = \frac{\int_0^h \boldsymbol{\Omega}_{(iN),t+\tau}\mathrm{d}\tau}{\int_0^h \boldsymbol{\Omega}_{(NN),t+\tau}\mathrm{d}\tau} \tag{7}$$

almost surely for all t as the sampling frequency increases, or $\Delta \to 0$.

A number of comments are in order. First, the integrated return covariance matrix, $\int_0^h \boldsymbol{\Omega}_{t+\tau}\mathrm{d}\tau$, is treated as stochastic, so both the integrated market variance and the integrated covariances of individual equity returns with the market over $[t,\ t+h]$ are ex ante, as of time t, unobserved and governed by a non-degenerate (and potentially unknown) distribution. Moreover, the covariance matrix will generally vary continuously and randomly over the entire interval, so the integrated covariance matrix should be interpreted as the average realized covariation among the return series. Second, Eq. (3) makes it clear that the realized market volatility in (4) and the realized covariance in (5) are continuous-record consistent estimators of the (random) realizations of the underlying integrated market volatility and covariance. Thus, as a corollary, the realized beta will be consistent for the integrated beta, as stated in (7). Third, the general representation here

encompasses the standard assumption of a constant beta over the measurement or estimation horizon, which is attained for the degenerate case of the Ω_t process being constant throughout each successive h-period measurement interval, or $\Omega_t = \Omega$. Fourth, the realized beta estimation procedure in Eq. (4)–(6) is implemented through a simple regression (without a constant term) of individual high-frequency stock returns on the corresponding market return. Nonetheless, the interpretation is very different from a standard regression, as the Ordinary Least Square (OLS) point estimate now represents a consistent estimator of the ex post realized regression coefficient obtained as the ratio of unbiased estimators of the average realized covariance and the realized market variance. The associated continuous-record asymptotic theory developed by Barndorff-Nielsen and Shephard (2003) explicitly recognizes the diffusion setting underlying this regression interpretation and hence facilitates the construction of standard errors for our beta estimators.

2.2. Nonlinear Fractional Cointegration: A Common Long-Memory Feature in Variances and Covariances

The possibility of common persistent components is widely recognized in modern multivariate time-series econometrics. It is also important for our analysis, because there may be common persistence features in the underlying variances and covariances from which betas are produced.

The idea of a common feature is a simple generalization of the well-known cointegration concept. If two variables are integrated but there exists a function f of them that is not, we say that they are cointegrated, and we call f the conintegrating function. More generally, if two variables have property X but there exists a function of them that does not, we say that they have common feature X. A key situation is when X corresponds to *persistence*, in which case we call the function of the two variables that eliminates the persistence the *copersistence function*. It will prove useful to consider linear and nonlinear copersistence functions in turn.

Most literature focuses on linear copersistence functions. The huge cointegration literature pioneered by Granger (1981) and Engle and Granger (1987) deals primarily with linear common long-memory $I(1)$ persistence features. The smaller copersistence literature started by Engle and Kozicki (1993) deals mostly with linear common short-memory $I(0)$ persistence features. The idea of fractional cointegration, suggested by Engle and Granger (1987) and developed by Cheung and Lai (1993) and Robinson and Marinucci, (2001),

among others, deals with linear common long-memory $I(d)$ persistence features, $0<d<1/2$.

Our interest is closely related but different. First, it centers on *nonlinear* copersistence functions, because betas are ratios. There is little literature on nonlinear common persistence features, although they are implicitly treated in Granger (1995). We will be interested in nonlinear common long-memory $I(d)$ persistence features, $0<d<1/2$, effectively corresponding to nonlinear fractional cointegration.[8]

Second, we are interested primarily in the case of *known* cointegrating relationships. That is, we may not know whether a given stock's covariance with the market is fractionally cointegrated with the market variance, but if it is, then there is a good financial economic reason (i.e., the CAPM) to suspect that the cointegrating function is the *ratio* of the covariance to the variance. This provides great simplification. In the integer-cointegration framework with known cointegrating vector under the alternative, for example, one could simply test the cointegrating combination for a unit root, or test the significance of the error-correction term in a complete error-correction model, as in Horvath and Watson (1995). We proceed in analogous fashion, examining the integration status (generalized to allow for fractional integration) of the realized market variance, realized individual equity covariances with the market, and realized market betas.

Our realized beta series are unfortunately relatively short compared to the length required for formal testing and inference procedures regarding (fractional) cointegration, as the fractional integration and cointegration estimators proposed by Geweke and Porter-Hudak (1983), Robinson and Marinucci (2001), and Andrews and Guggenberger (2003) tend to behave quite erratically in small samples. In addition, there is considerable measurement noise in the individual beta series so that influential outliers may have a detrimental impact on our ability to discern the underlying dynamics. Hence, we study the nature of the long range dependence and short-run dynamics in the realized volatility measures and realized betas through intentionally less formal but arguably more informative graphical means, and via some robust procedures that utilize the joint information across many series, to which we now turn.

3. EMPIRICAL ANALYSIS

We examine primarily the realized quarterly betas constructed from daily returns. We focus on the dynamic properties of market betas vis-à-vis the

dynamic properties of their underlying covariance and variance components. We quantify the dynamics in a number of ways, including explicit measurement of the degree of predictability in the tradition of Granger and Newbold (1986).

3.1. Dynamics of Quarterly Realized Variance, Covariances and Betas

This section investigates the realized quarterly betas constructed from daily returns obtained from the Center for Research in Security Prices from July 1962 to September 1999. We take the market return $r_{m,t}$ to be the 30 Dow Jones Industrial Average (DJIA), and we study the subset of 25 DJIA stocks as of March 1997 with complete data from July 2, 1962 to September 17, 1999, as detailed in Table 1. We then construct quarterly realized DJIA variances, individual equity covariances with the market, and betas, 1962:3–1999:3 (149 observations).

In Fig. 1, we provide a time-series plot of the quarterly realized market variance, with fall 1987 included (top panel) and excluded (bottom panel). It is clear that the realized variance is quite persistent and, moreover, that the fall 1987 volatility shock is unlike any other ever recorded, in that volatility reverts to its mean almost instantaneously. In addition, our subsequent computation of asymptotic standard errors reveals that the uncertainty associated with the fall 1987 beta estimate is enormous, to the point of rendering it entirely uninformative. In sum, it is an exceptional outlier with potentially large influence on the analysis, and it is measured with huge imprecision. Hence, following many other authors, we drop the fall 1987 observation from this point onward.

In Figs. 2 and 3, we display time-series plots of the 25 quarterly realized covariances and realized betas.[9] Like the realized variance, the realized covariances appear highly persistent. The realized betas, in contrast, appear noticeably less persistent. This impression is confirmed by the statistics presented in Table 2: the mean Ljung–Box Q-statistic (through displacement 12) is 84 for the realized covariance, but only 47 for the realized beta, although both are of course significant relative to a $\chi^2(12)$ distribution.[10]

The impression of reduced persistence in realized betas relative to realized covariances is also confirmed by the sample autocorrelation functions for the realized market variance, the realized covariances with the market, and the realized betas shown in Fig. 4.[11] Most remarkable is the close correspondence between the shape of the realized market variance correlogram and the realized covariance correlograms. This reflects an extraordinary high degree of dependence in the correlograms across the individual realized

Table 1. The Dow Jones Thirty.

Company Name	Ticker	Data Range
Alcoa Inc.	AA	07/02/1962 to 09/17/1999
Allied Capital Corporation	ALD	07/02/1962 to 09/17/1999
American Express Co.	AXP[a]	05/31/1977 to 09/17/1999
Boeing Co.	BA	07/02/1962 to 09/17/1999
Caterpillar Inc.	CAT	07/02/1962 to 09/17/1999
Chevron Corp.	CHV	07/02/1962 to 09/17/1999
DuPont Co.	DD	07/02/1962 to 09/17/1999
Walt Disney Co.	DIS	07/02/1962 to 09/17/1999
Eastman Kodak Co.	EK	07/02/1962 to 09/17/1999
General Electric Co.	GE	07/02/1962 to 09/17/1999
General Motors Corp.	GM	07/02/1962 to 09/17/1999
Goodyear Tire & Rubber Co.	GT	07/02/1962 to 09/17/1999
Hewlett–Packard Co.	HWP	07/02/1962 to 09/17/1999
International Business Machines Corp.	IBM	07/02/1962 to 09/17/1999
International Paper Co.	IP	07/02/1962 to 09/17/1999
Johnson & Johnson	JNJ	07/02/1962 to 09/17/1999
JP Morgan Chase & Co.	JPM[a]	03/05/1969 to 09/17/1999
Coca–Cola Co.	KO	07/02/1962 to 09/17/1999
McDonald's Corp.	MCD[a]	07/05/1966 to 09/17/1999
Minnesota Mining & Manufacturing Co.	MMM	07/02/1962 to 09/17/1999
Philip Morris Co.	MO	07/02/1962 to 09/17/1999
Merck & Co.	MRK	07/02/1962 to 09/17/1999
Procter & Gamble Co.	PG	07/02/1962 to 09/17/1999
Sears, Roebuck and Co.	S	07/02/1962 to 09/17/1999
AT&T Corp.	T	07/02/1962 to 09/17/1999
Travelers Group Inc.	TRV[a]	10/29/1986 to 09/17/1999
Union Carbide Corp.	UK	07/02/1962 to 09/17/1999
United Technologies Corp.	UTX	07/02/1962 to 09/17/1999
Wal–Mart Stores Inc.	WMT[a]	11/20/1972 to 09/17/1999
Exxon Corp.	XON	07/02/1962 to 09/17/1999

Note: A summary of company names and tickers, and the range of the data are examined. We use the Dow Jones Thirty as of March 1997.
[a]Stocks with incomplete data, which we exclude from the analysis.

covariances with the market, as shown in Fig. 5. In Fig. 4, it makes the median covariance correlogram appear as a very slightly dampened version of that for the market variance. This contrasts sharply with the lower and gently declining pattern for the realized beta autocorrelations. Intuitively, movements of the realized market variance are largely reflected in movements of the realized covariances; as such, they largely "cancel" when we form ratios (realized betas). Consequently, the correlation structure across

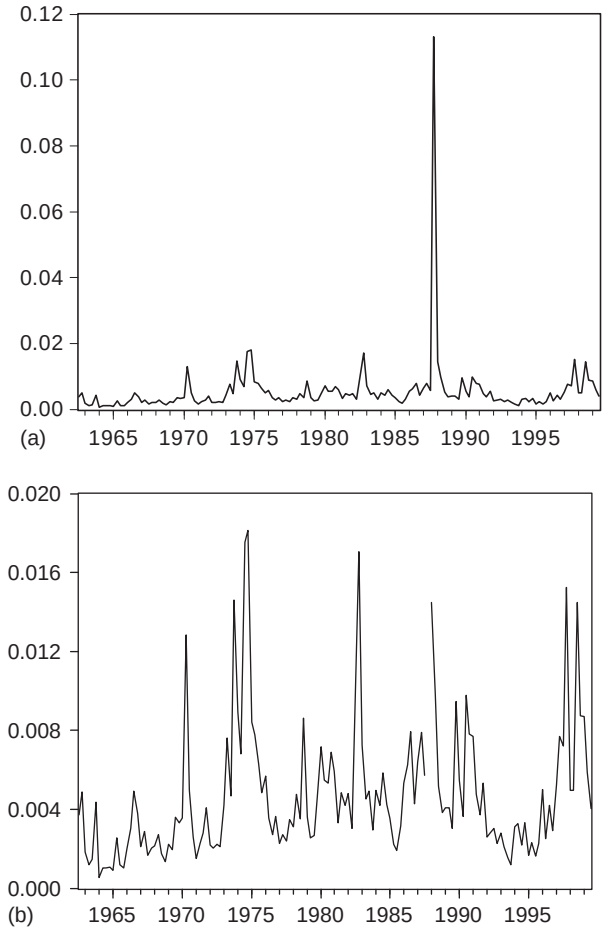


Fig. 1. Time Series Plot of Quarterly Realized Market Variance, Fall 1987 (a) Included (b) Exculded. *Note:* The Two Subfigures Show the Time Series of Quarterly Realized Market Variance, with The 1987:4 Outlier Included (a) and Excluded (b). The Sample Covers the Period from 1962:3 through 1999:3, for a Total of 149 Observations. We Calculate the Realized Quarterly Market Variances from Daily Returns.

the individual realized beta series in Fig. 6 is much more dispersed than is the case for the realized covariances in Fig. 5. This results in an effective averaging of the noise and the point estimates of the median correlation values are effectively zero beyond 10 quarters for the beta series.[12]

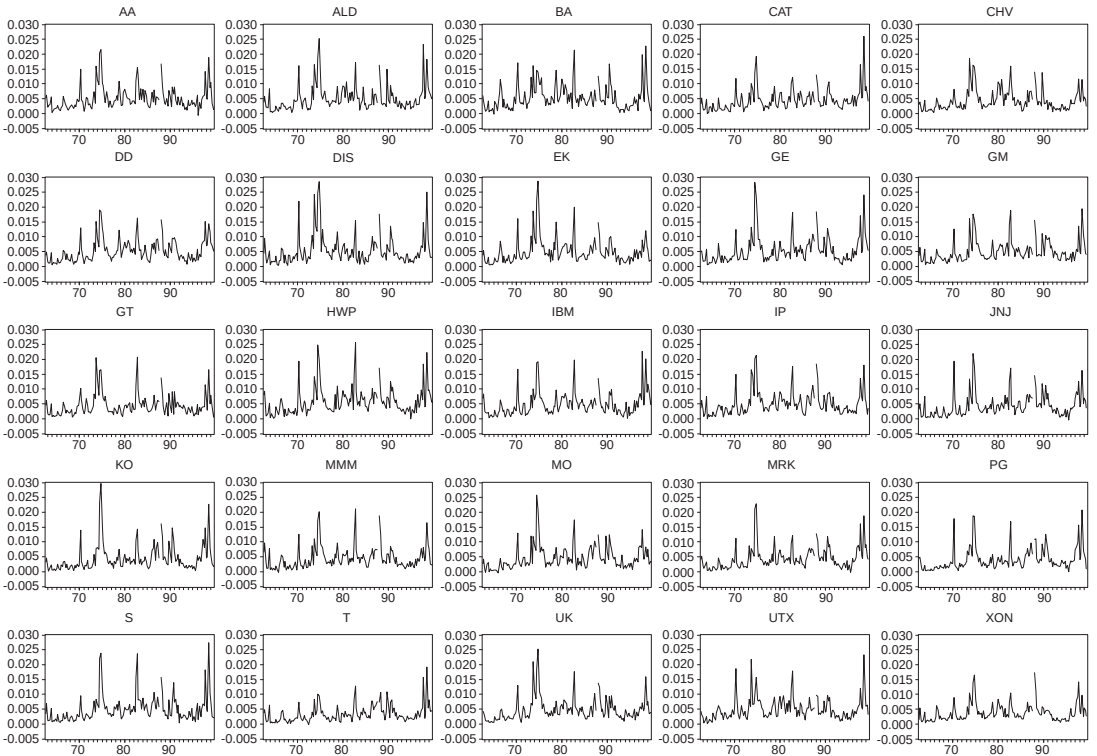


Fig. 2. Time Series Plots of Quarterly Realized Covariances. *Note*: The Time Series of Quarterly Realized Covariances, with The 1987:4 Outlier Excluded are Shown. The Sample Covers The Period from 1962:3 through 1999:3, for a Total of 148 Observations. We Calculate the Realized Quarterly Covariances from Daily Returns.

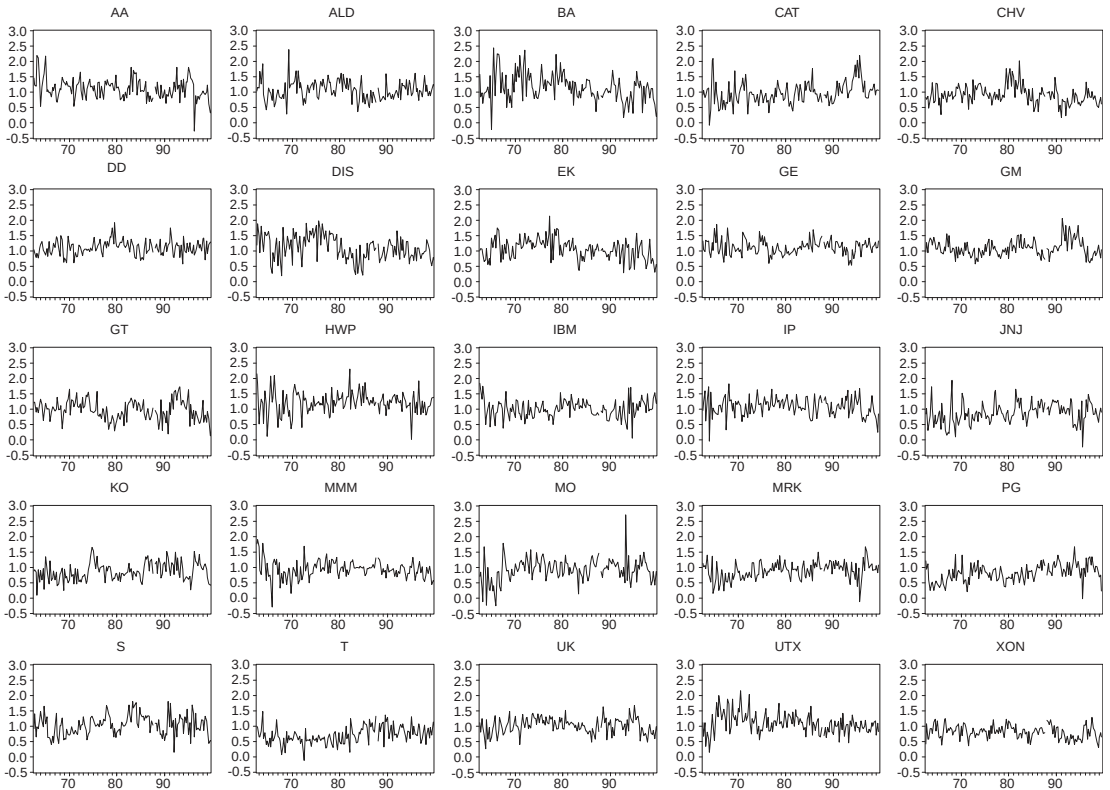


Fig. 3. Time Series Plots of Quarterly Realized Betas. *Note:* The Time Series of Quarterly Realized Betas, with The 1987:4 Outlier Excluded are Shown. The Sample Covers The Period from 1962:3 through 1999:3, for a Total of 148 Observations. We Calculate the Realized Quarterly Betas from Daily Returns.

Table 2. The Dynamics of Quarterly Realized Market Variance, Covariances and Betas.

v_{mt}^2	Q	ADF^1	ADF^2	ADF^3	ADF^4
	1.09.50	−5.159	−3.792	−4.014	−3.428

	$cov(\mathbf{r}_{mt}, \mathbf{r}_{it})$					β_{it}				
	Q	ADF^1	ADF^2	ADF^3	ADF^4	Q	ADF^1	ADF^2	ADF^3	ADF^4
Min.	47.765	−6.188	−4.651	−4.621	−4.023	6.6340	−8.658	−6.750	−5.482	−5.252
0.10	58.095	−5.880	−4.383	−4.469	−3.834	15.026	−7.445	−6.419	−5.426	−4.877
0.25	69.948	−5.692	−4.239	−4.352	−3.742	26.267	−6.425	−5.576	−5.047	−4.294
0.50	84.190	−5.478	−4.078	−4.179	−3.631	46.593	−6.124	−5.026	−3.896	−3.728
0.75	100.19	−5.235	−3.979	−4.003	−3.438	66.842	−5.431	−4.188	−3.724	−3.313
0.90	119.28	−4.915	−3.777	−3.738	−3.253	106.67	−4.701	−3.404	−3.225	−2.980
Max.	150.96	−4.499	−3.356	−3.690	−2.986	134.71	−4.600	−3.315	−2.808	−2.493
Mean	87.044	−5.435	−4.085	−4.159	−3.580	53.771	−6.090	−4.925	−4.245	−3.838
S.D.	24.507	−0.386	0.272	0.250	0.239	35.780	1.026	0.999	0.802	0.729

Note: Summary on the aspects of the time-series dependence structure of quarterly realized market variance, covariances, and realized betas. Q denotes the Ljung–Box portmanteau statistic for up to 12th-order autocorrelation, and ADF^i denotes the augmented Dickey–Fuller unit root test, with ntercept and with i augmentation lags. The sample covers the period from 1962:3 through 1999:3, with the 1987:4 outlier excluded, for a total of 148 observations. We calculate the quarterly realized variance, covariances, and betas from daily returns.

The work of Andersen et al. (2001a) and Andersen et al. (2003), as well as that of many other authors, indicates that asset return volatilities are well-described by a pure fractional noise process, typically with the degree of integration around $d \approx 0.4$.[13] That style of analysis is mostly conducted on high-frequency data. Very little work has been done on long memory in equity variances, market covariances, and market betas at the quarterly frequency, and it is hard to squeeze accurate information about d directly from the fairly limited quarterly sample. It is well-known, however, that if a flow variable is $I(d)$, then it remains $I(d)$ under temporal aggregation. Hence, we can use the results of analyses of high-frequency data, such as Andersen et al. (2003), to help us analyze the quarterly data. After some experimentation, and in keeping with the typical finding that $d \approx 0.4$, we settled on $d = 0.42$.

In Fig. 7, we graph the sample autocorrelations of the quarterly realized market variance, the median realized covariances with the market, and the median realized betas, all prefiltered by $(1-L)^{0.42}$. It is evident that the dynamics in the realized variance and covariances are effectively annihilated by filtering with $(1-L)^{0.42}$, indicating that the pure fractional noise process with $d = 0.42$ is indeed a good approximation to their dynamics. Interestingly,

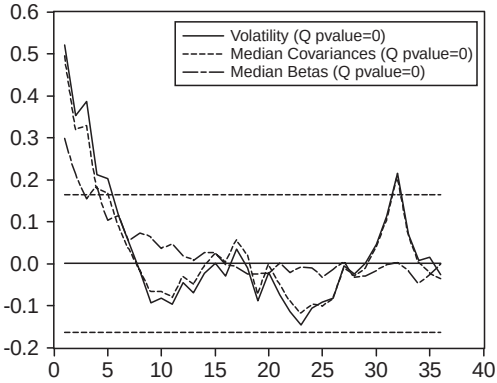


Fig. 4. Sample Autocorrelations of Quarterly Realized Market Variance, Median Sample Autocorrelations of Quarterly Realized Covariances and Median Sample Autocorrelations of Quarterly Realized Betas. *Note:* The First 36 Sample Autocorrelations of Quarterly Realized Market Variance, the Medians across Individual Stocks of the First 36 Sample Autocorrelations of Quarterly Realized Covariances and the Medians across Individual Stocks of the First 36 Sample Autocorrelations of Quarterly Realized Betas are Shown. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White Noise Case. Q Denotes the Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Variance, Covariances, and Betas from Daily Returns.

however, filtering the realized *betas* with $(1-L)^{0.42}$ appears to produce *over-differencing*, as evidenced by the fact that the first autocorrelation of the fractionally differenced betas is often negative. Compare, in particular, the median sample autocorrelation function for the prefiltered realized covariances to the median sample autocorrelation function for the prefiltered realized betas. The difference is striking in the sense that the first autocorrelation coefficient for the betas is negative and much larger than those for all of the subsequent lags. Recall that the standard error band for the median realized beta (not shown in the lower panels, as it depends on the unknown cross-sectional dependence structure) should be considerably narrower than for the other series in Fig. 7, thus likely rendering the first-order correlation coefficient for the beta series significantly negative. This finding can be seen to be reasonably consistent across the individual prefiltered covariance and beta correlation functions displayed in Figs. 8 and 9.

If fractional differencing of the realized betas by $(1-L)^{0.42}$ may be "too much," then the question naturally arises as to how much differencing is "just

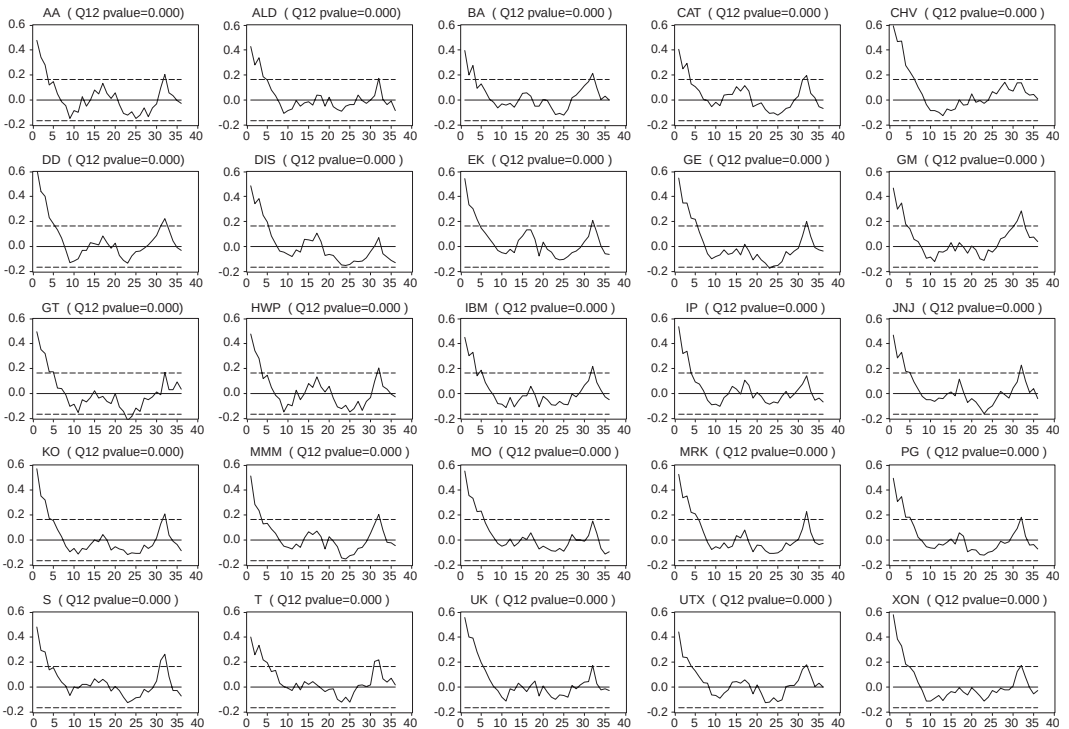


Fig. 5. Sample Autocorrelations of Quarterly Realized Covariances. *Note:* The Figure Shows the First 36 Sample Autocorrelations of Quarterly Realized Covariances are Shown. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White Noise Case. Q Denotes the Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Covariances from Daily Returns.

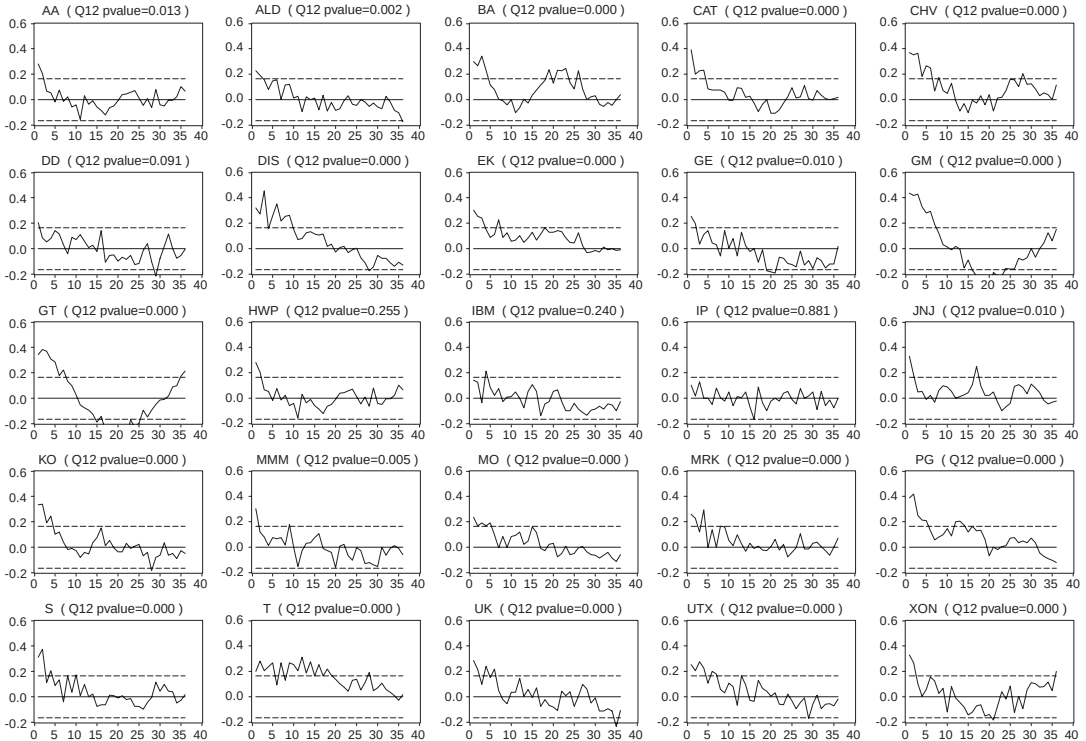


Fig. 6. Sample Autocorrelations of Quarterly Realized Betas. *Note:* The First 36 Sample Autocorrelations of Quarterly Realized Betas are Shown. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White-Noise Case. Q Denotes the Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Betas from Daily Returns.

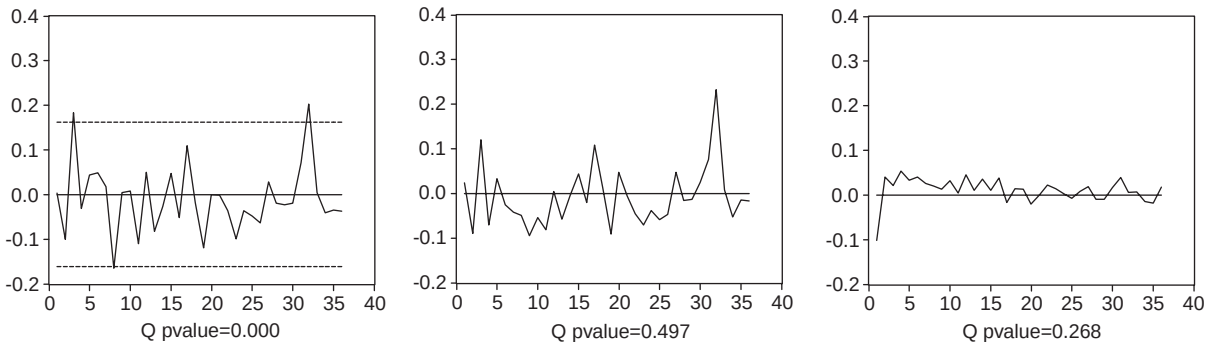


Fig. 7. (a) Sample Autocorrelations of Quarterly Realized Market Variance Prefiltered by $(1-L)^{0.42}$. Median Sample Autocorrelations of Quarterly Realized (b) Covariances and (c) Betas. *Note:* The Three Parts (a–c) Show the First 36 Sample Autocorrelations of Quarterly Realized Market Variance, the Medians across Individual Stocks of First 36 Sample Autocorrelations of Quarterly Realized Covariances and the Medians across Individual Stocks of First 36 Sample Autocorrelations of Quarterly Realized Betas all Prefiltered by $(1-L)^{0.42}$. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White-Noise Case. Q Denotes the Median of Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Variance from Daily Returns.

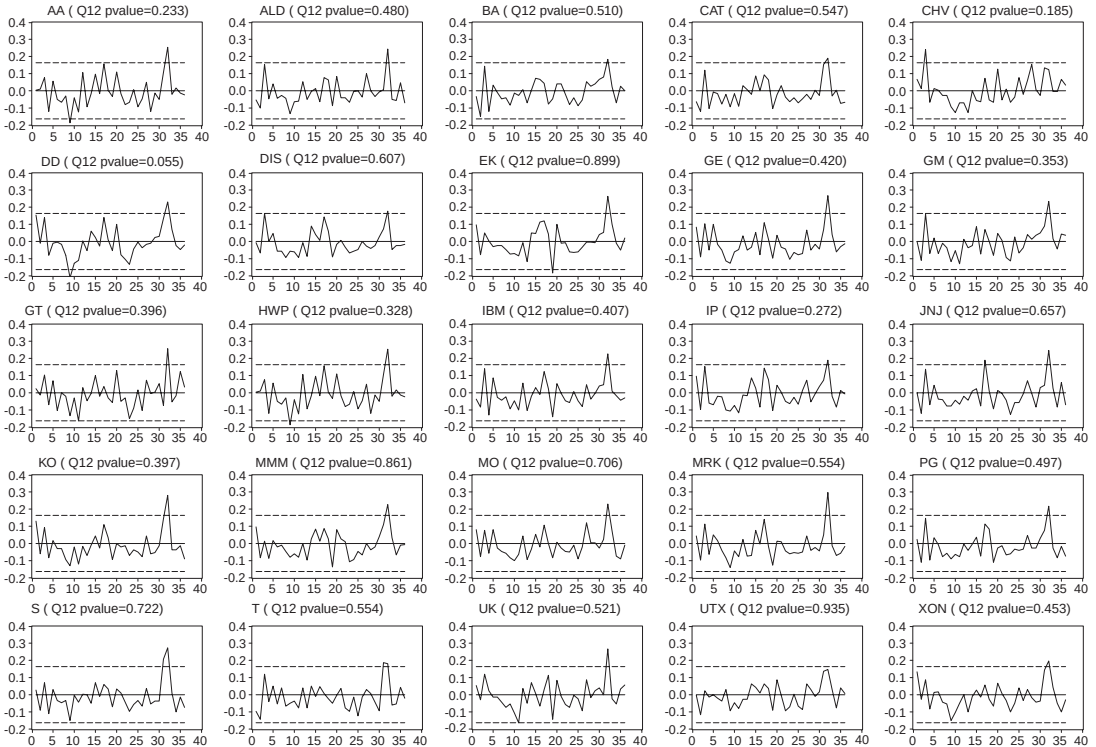


Fig. 8. Sample Autocorrelations of Quarterly Realized Covariances Prefiltered by $(1-L)^{0.42}$. *Note:* The First 36 Sample Autocorrelations of Quarterly Realized Covariances Prefiltered by $(1-L)^{0.42}$ is Shown. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White-Noise Case. Q Denotes the Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Covariances from Daily Returns.

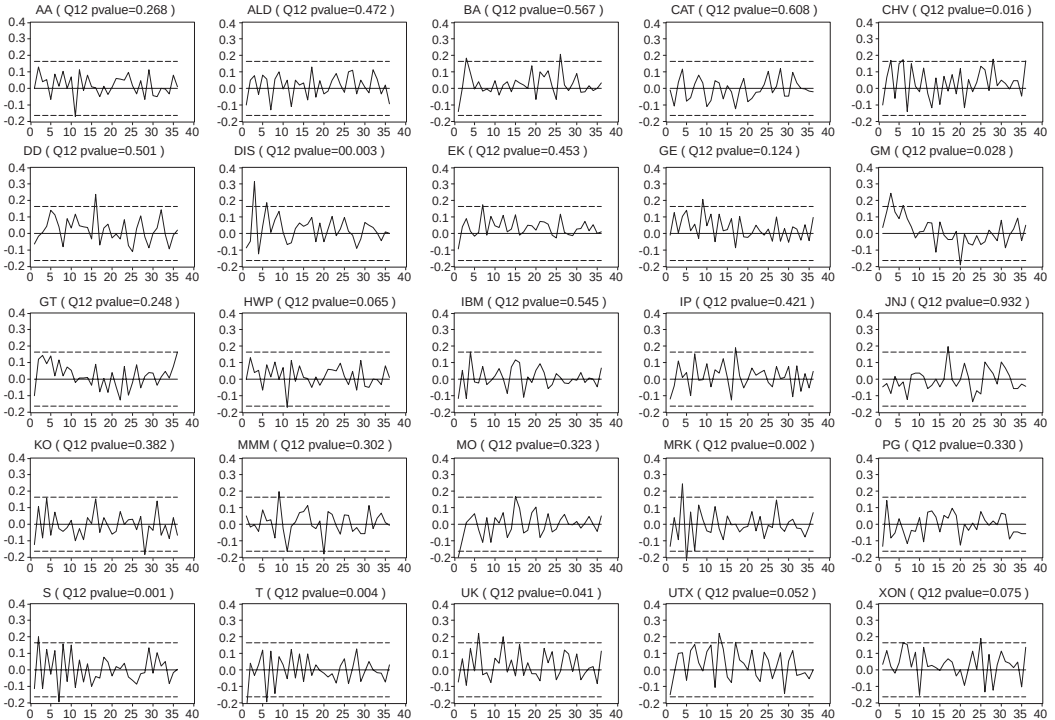


Fig. 9. Sample Autocorrelations of Quarterly Realized Betas Prefiltered by $(1-L)^{0.42}$. *Note:* The First 36 Sample Autocorrelations of Quarterly Realized Betas Prefiltered by $(1-L)^{0.42}$ are Shown. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White-Noise Case. Q Denotes the Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Betas from Daily Returns.

right." Some experimentation revealed that differencing the betas by $(1-L)^{0.20}$ was often adequate for eliminating the dynamics. However, for short samples it is almost impossible to distinguish low-order fractional integration from persistent but strictly stationary dynamics. We are particularly interested in the latter alternative where the realized betas are $I(0)$. To explore this possibility, we fit simple $AR(p)$ processes to realized betas, with p selected by the Akaike Information Criterion (AIC). We show the estimated roots in Table 3, all of which are indicative of covariance stationarity. In Fig. 10, we show the sample autocorrelation functions of quarterly realized betas prefiltered by the estimated $AR(p)$ lag-operator polynomials. The autocorrelation functions are indistinguishable from those of white noise.

Taken as a whole, the results suggest that realized betas are integrated of noticeably lower order than are the market variance and the individual equity covariances with the market, corresponding to a situation of nonlinear fractional cointegration. $I(d)$ behavior, with $d \in [0, 0.25]$, appears accurate for betas, whereas the market variance and the individual equity covariances with the market are better approximated as $I(d)$ with $d \in [0.35, 0.45]$. Indeed, there is little evidence against an assertion that betas are $I(0)$, whereas there is strong evidence against such an assertion for the variance and covariance components.

3.2. Predictability

Examination of the *predictability* of realized beta and its components provides a complementary perspective and additional insight. Granger and Newbold (1986) propose a measure of the predictability of covariance stationary series under squared-error loss, patterned after the familiar regression R^2,

$$G(j) = \frac{var(\hat{x}_{t+j,t})}{var(x_t)} = 1 - \frac{var(e_{t+j,t})}{var(x_t)}, \tag{8}$$

where j is the forecast horizon of interest, $\hat{x}_{t+j,t}$ the optimal (i.e., conditional mean) forecast, and $e_{t+j,t} = x_{t+j} - \hat{x}_{t+j,t}$. Diebold and Kilian (2001) define a generalized measure of predictability, building on the Granger–Newbold measure, as

$$P(L, \Omega, j, k) = 1 - \frac{E(L(e_{t+j,t}))}{E(L(e_{t+k,t}))}, \tag{9}$$

where L denotes the relevant loss function, Ω is the available univariate or multivariate information set, j the forecast horizon of interest, and k a long but not necessarily infinite reference horizon.

Table 3. Inverted Roots of $AR(p)$ Models for Quarterly Realized Betas.

Stock	Inverted Roots					(and Modulus of Dominant Inverted Root)
AA	0.49−0.25i	0.49+0.25i	−0.10−0.41i	−0.10+0.41i	−0.57	(0.57)
ALD	0.50	−0.30				(0.50)
BA	0.80	−0.30+0.49i	−0.30−0.49i			(0.80)
CAT	0.39					(0.39)
CHV	0.80	−0.29−0.44i	−0.29+0.44i			(0.80)
DD	0.20					(0.20)
DIS	0.86	0.20−0.48i	0.20+0.48i	−0.50−0.59i	−0.50+0.59i	(0.86)
EK	0.73	−0.25+0.38i	−0.25−0.38i			(0.73)
GE	0.50	−0.28				(0.50)
GM	0.84	−0.29+0.44i	−0.29−0.44i			(0.84)
GT	0.83	−0.33+0.41i	−0.33−0.41i			(0.83)
HWP	0.36	−0.13+0.27i	−0.13−0.27i			(0.36)
IBM	0.66	0.09+0.68i	0.09−0.68i	−0.76		(0.76)
IP	0.10					(0.10)
JNJ	0.33					(0.33)
KO	0.79	0.04+0.50i	0.04−0.50i	−0.63		(0.79)
MMM	0.47	−0.13+0.31i	−0.13−0.31i			(0.47)
MO	0.83	0.16+0.61i	0.16−0.61i	−0.48−0.35i	−0.48+0.35i	(0.83)
MRK	0.60−0.11i	0.60+0.11i	−0.07−0.73i	−0.07+0.73i	−0.81	(0.81)
PG	0.72	−0.45				(0.72)
S	0.59	0.25	−0.57			(0.59)
T	0.87	0.17−0.64i	0.17+0.64i	−0.56+0.34i	−0.56−0.34i	(0.87)
UTX	0.77	0.08+0.63i	0.08−0.63i	−0.68		(0.77)
UK	0.80	−0.10+0.58i	−0.10−0.58i	−0.42		(0.80)
XON	0.58	−0.29				(0.58)

Note: The inverted roots and modulus of the dominant root of the autoregressive lag operator polynomials $(1 - \hat{\Phi}_1 L - \hat{\Phi}_2 L^2 \ldots - \hat{\Phi}_p L^p)$, where $\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_p$ are the least-squares estimates of the parameters of $AR(p)$ models fit to the realized betas, with p selected by the AIC are shown. The sample covers the period from 1962:3 through 1999:3, with the 1987:4 outlier excluded, for a total of 148 observations. We calculate the quarterly realized variance, covariances, and betas from daily returns.

Regardless of the details, the basic idea of predictability measurement is simply to compare the expected loss of a short-horizon forecast to the expected loss of a very long-horizon forecast. The former will be much smaller than the latter if the series is highly predictable, as the available conditioning information will then be very valuable. The Granger–Newbold measure, which is the canonical case of the Diebold–Kilian measure (corresponding to $L(e) = e^2$, univariate Ω, and $k = \infty$) compares the 1-step-ahead forecast error variance to that of the ∞-step-ahead forecast error variance, i.e., the unconditional variance of the series being forecast (assuming that it is finite).

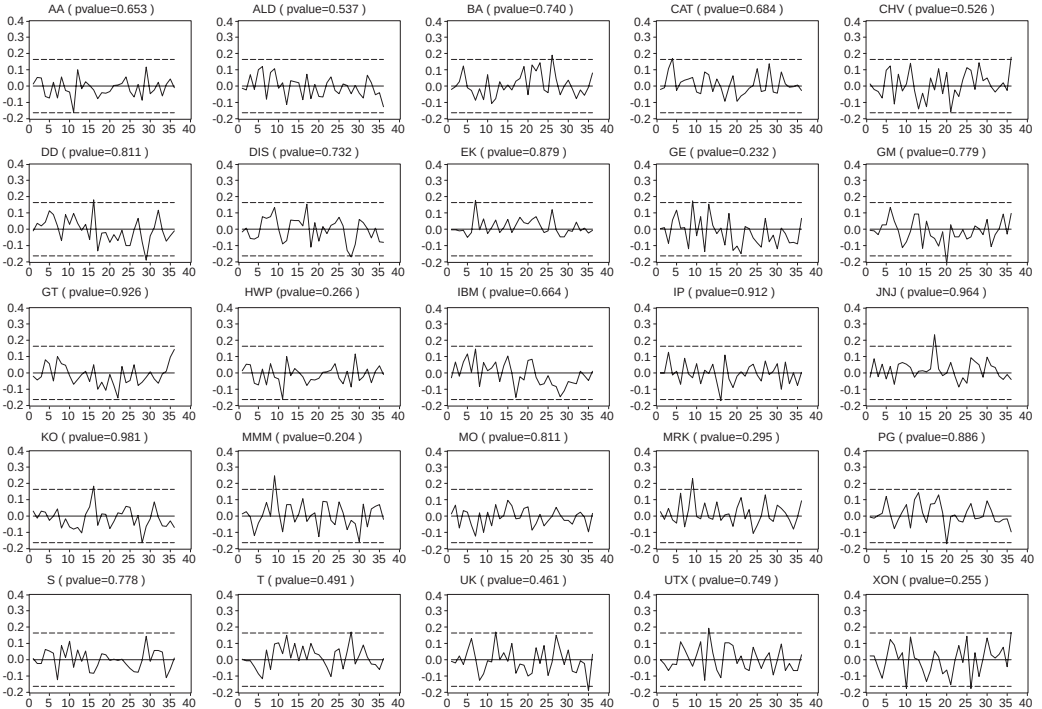


Fig. 10. Sample Autocorrelations of Quarterly Realized Betas Prefiltered by $(1 - \hat{\Phi}_1 L - \hat{\Phi}_2 L^2 - \cdots - \hat{\Phi}_p L^p)$. *Note:* The First 36 Sample Autocorrelations of Quarterly Realized Betas Prefiltered by $(1 - \hat{\Phi}_1 L - \hat{\Phi}_2 L^2 - \cdots - \hat{\Phi}_p L^p)$, where $\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_p$ are the Least Squares Estimates of the Parameters of $AR(p)$ Models Fit to the Realized betas, with p Selected by the AIC are Shown. The Dashed Lines Denote Bartlett's Approximate 95 Percent Confidence Band in the White-Noise Case. Q Denotes the Ljung–Box Portmanteau Statistic for up to 12th-Order Autocorrelation. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Variance, Covariances, and Betas from Daily Returns.

In what follows, we use predictability measures to provide additional insight into the comparative dynamics of the realized variances and covariances versus the realized betas. Given the strong evidence of fractional integration in the realized market variance and covariances, we maintain the pure fractional noise process for the quarterly realized market variance and the realized covariances, namely $ARFIMA(0, 0.42, 0)$. We then calculate the Granger–Newbold predictability $G(j)$ analytically, conditional upon the $ARFIMA(0, 0.42, 0)$ dynamics, and we graph it in Fig. 11 for $j = 1,\ldots, 7$ quarters.[14] The graph starts out as high as 0.4 and decays only slowly over the first seven quarters. If the realized beta likewise follows a pure fractional noise process but with a smaller degree of integration, say $ARFIMA(0, 0.20, 0)$, which we argued was plausible, then the implied predictability is much lower, as also shown in Fig. 11. As we also argued, however, the integration status of the realized betas is difficult to determine. Hence, for the realized betas we also compute Granger–Newbold predictability using an estimated $AR(p)$ sieve approximation to produce estimates of $var(e_{t+j,t})$ and $var(x_t)$; this approach is valid regardless of whether the true dynamics are short-memory or long-memory. In Fig. 12 we plot the beta predictabilities, which remain noticeably smaller and more quickly decaying than the covariance predictabilities, as is further clarified by comparing the median beta predictability, also included in Fig. 11, to the market variance and equity covariances predictability. It is noteworthy that the shorter-run beta predictability – up to about four quarters – implied by the $AR(p)$ dynamics is considerably higher than for the $I(0.20)$ dynamics. Due to the long-memory feature of the $I(0.20)$ process this eventually reverses beyond five quarters.

4. ASSESSING PRECISION: INTERVAL ESTIMATES OF BETAS

Thus far, we have largely abstracted from the presence of estimation error in the realized betas. It is possible to assess the (time-varying) estimation error directly using formal continuous-record asymptotics.

4.1. Continuous-Record Asymptotic Standard Errors

We first use the multivariate asymptotic theory recently developed by Barndorff-Nielsen and Shephard (2003) to assess the precision of our realized betas which are, of course, estimates of the underlying integrated

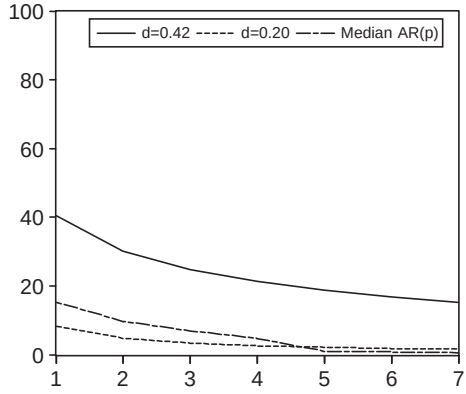


Fig. 11. Predictability of Market Volatility, Individual Equity Covariances with the Market, and Betas. *Note:* We Define Predictability as $P_j = 1 - var(e_{t+j,t})/var(e_{t+40,t})$, where $var(e_{t+j,t}) = \sigma^2 \sum_{i=0}^{j-1} b_i^2$, σ_t^2 is the Variance of the Innovation ε_t, and the b_i's are Moving Average Coefficients; i.e., the Wold Representation is $y_t = (1 + b_1 L + b_2 L^2 + b_2 L^3 + \ldots)\varepsilon_t$. We Approximate the Dynamics Using a Pure Long-Memory Model, $(1-L)^{0.42}\, y_t = \varepsilon_t$, in Which Case $b_0 = 1$ and $b_i = (-1)b_{i-1}(d - i + 2)/(i - 1)$ and Plot P_j for $j = 1,\ldots, 7$ in the Solid Line. Moreover, Because We Take $d = 0.42$ for Market Volatility and for all Covariances with the Market, all of Their Predictabilities are the Same at all Horizons. As One Approximation for the Dynamics of the Betas we use a Pure Long-Memory Model, $(1-L)^{0.20}\, y_t = \varepsilon_t$, in Which Case $b_0 = 1$ and $b_i = (-1)b_{i-1}(d - i + 2)/(i - 1)$ and Plot P_j for $j = 1,\ldots, 7$ in the Dotted Line. We also Approximate the Beta Dynamics Using an $AR(p)$ Model, with the Autoregressive Lag Order p Determined by the AIC and Plot the Median of P_j for $j = 1,\ldots, 7$ among all 25 Stocks in the Mixed Dotted Line. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations.

betas. This helps us in thinking about separating "news from noise" when examining temporal movements in the series.

From the discussion above, realized beta for stock i in quarter t is simply

$$\hat{\beta}_{it} = \frac{\sum_{j=1}^{N_t} \boldsymbol{r}_{ijt}\boldsymbol{r}_{mjt}}{\sum_{j=1}^{N_t} \boldsymbol{r}_{mjt}^2}, \tag{10}$$

where r_{ijt} is the return of stock i on day j of quarter t, $\boldsymbol{r}_{mjt}$ the return of the DJIA on day j of quarter t, and N_t the number of units (e.g., days) into which quarter t is partitioned.[15] Under appropriate regularity conditions that allow for non-stationarity in the series, Barndorff-Nielsen and Shephard (2003)

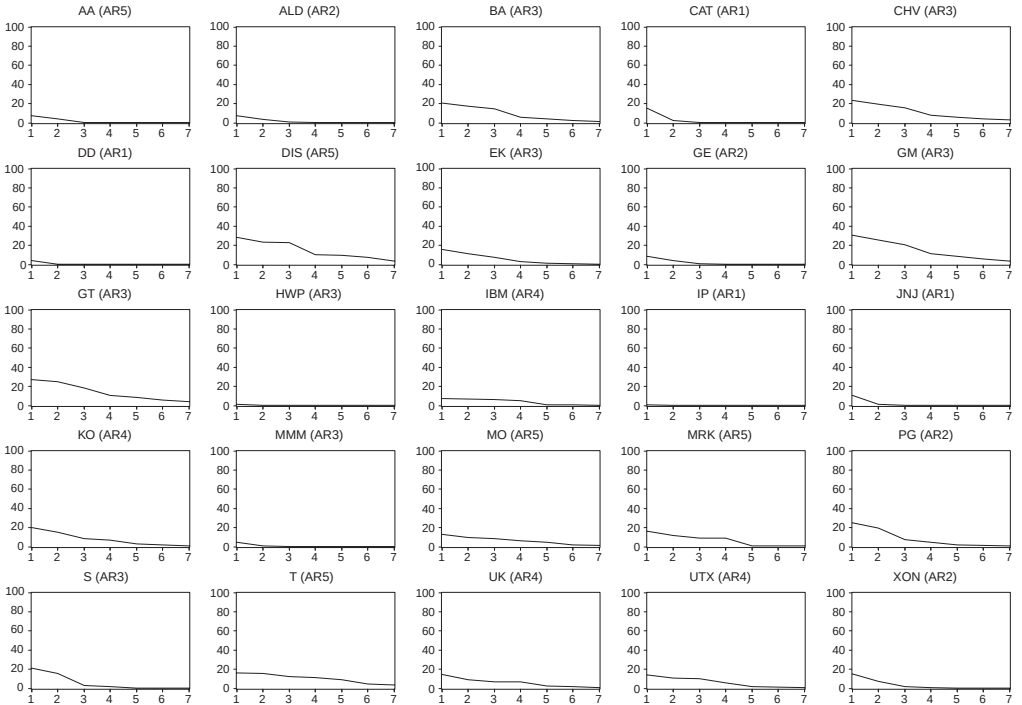


Fig. 12. Predictability of Betas Based on $AR(p)$ Sieve Approximation of Dynamics. *Note:* We Define Predictability as $P_j = 1 - var(e_{t+j,t})/var(y_t)$, Where $var(e_{t+j,t}) = \sigma^2 \sum_{i=0}^{j-1} b_i^2$, $var(y_t) = \sigma^2 \sum_{i=0}^{\infty} b_i^2$, σ_t^2 is the Variance of the Innovation ε_t, so that the b_i's Correspond to the Moving Average Coefficients in the Wold Representation for y_t. We Approximate the Dynamics Using an $AR(p)$ Model, with the Autoregressive Lag Order p Determined by the AIC, and Plot P_j for $j = 1, \ldots, 7$. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded, for a Total of 148 Observations. We Calculate the Quarterly Realized Betas from Daily Returns.

derive the limiting distribution of realized beta. In particular, as $N \to \infty$,

$$\frac{\hat{\beta}_{it} - \beta_{it}}{\sqrt{\left(\sum_{j=1}^{N_t} \boldsymbol{r}_{mjt}^2\right)^{-2} \hat{g}_{it}}} \Rightarrow N(0,1), \tag{11}$$

where

$$\hat{g}_{it} = \sum_{j=1}^{N_t} a_{ij}^2 - \sum_{j=1}^{N_t - 1} a_{ij} a_{ij+1} \tag{12}$$

and

$$a_{ij} = \boldsymbol{r}_{ijt} \boldsymbol{r}_{mjt} - \hat{\beta}_{it} \boldsymbol{r}_{mjt}^2. \tag{13}$$

Thus, a feasible and asymptotically valid α-percent confidence interval for the underlying integrated beta is

$$\beta_{it} \in \hat{\beta}_{it} \pm z_{\alpha/2} \sqrt{\left(\sum_{j=1}^{N_t} \boldsymbol{r}_{mjt}^2\right)^{-2} \hat{g}_{it}}, \tag{14}$$

where $z_{\alpha/2}$ denotes the appropriate critical value of the standard normal distribution.

In Fig. 13, we plot the pointwise 95 percent confidence intervals for the quarterly betas. They are quite wide, indicating that daily sampling is not adequate to drive out all measurement error. They are, given the width of the bands, moreover, consistent with the conjecture that there is only limited (short range) dependence in the realized beta series.

The continuous record asymptotics discussed above directly points to the advantage of using finer sampled data for improved beta measurements. However, the advent of reliable high-frequency intraday data is, unfortunately, a relatively recent phenomenon and we do not have access to such data for the full 1962:3–1999:3 sample period used in the empirical analysis so far. Nonetheless, to see how the reduction in measurement error afforded by the use of finer sample intradaily data manifests itself empirically in more reliable inference, we reproduce in Fig. 14 the pointwise 95 percent confidence bands for the quarterly betas over the shorter 1993:1–1999:3 sample. These bands may be compared directly to the corresponding quarterly realized beta standard error bands over the identical time span based on a 15-min sampling scheme reported in Fig. 15.[16] The improvement is readily visible in the narrowing of the bands. It is also evident from Fig. 15 that there is quite pronounced positive dependence in the realized quarterly beta measures. In

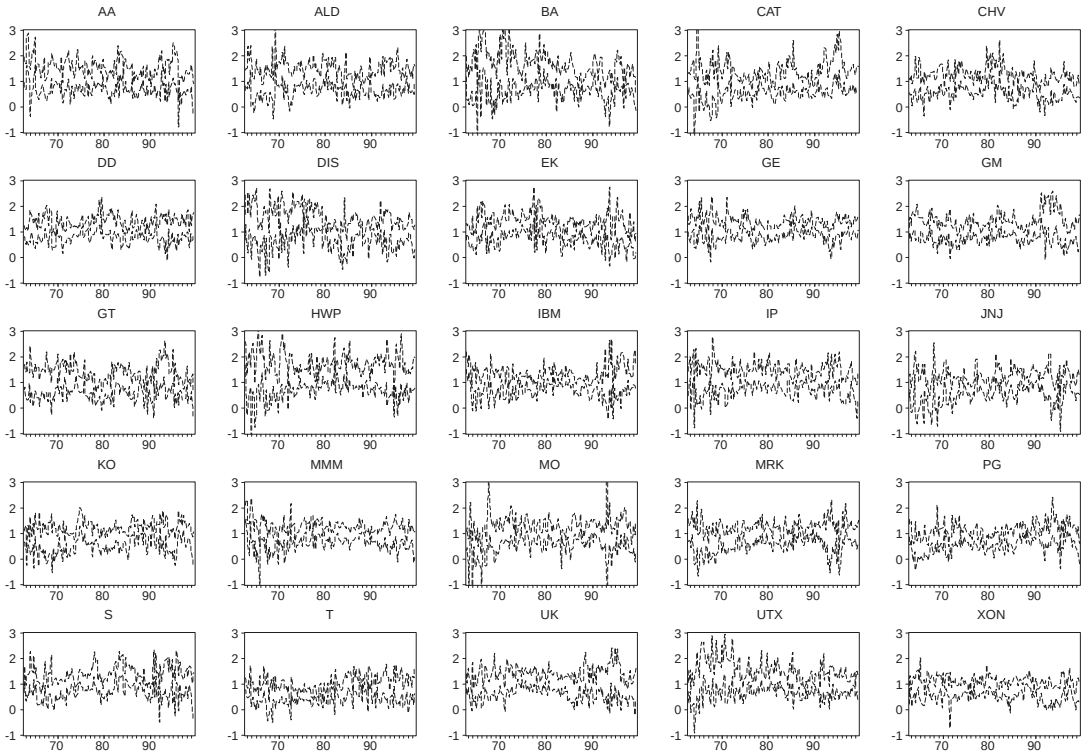


Fig. 13. Ninety-Five Percent Confidence Intervals for Quarterly Beta, Long Sample, Daily Sampling. *Note:* The Time Series of 95 Percent Confidence Intervals for the Underlying Quarterly Integrated Beta, Calculated Using the Results of Barndorff-Nielsen and Shephard (2003) are Shown. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded. We Calculate the Realized Quarterly Betas from Daily Returns.

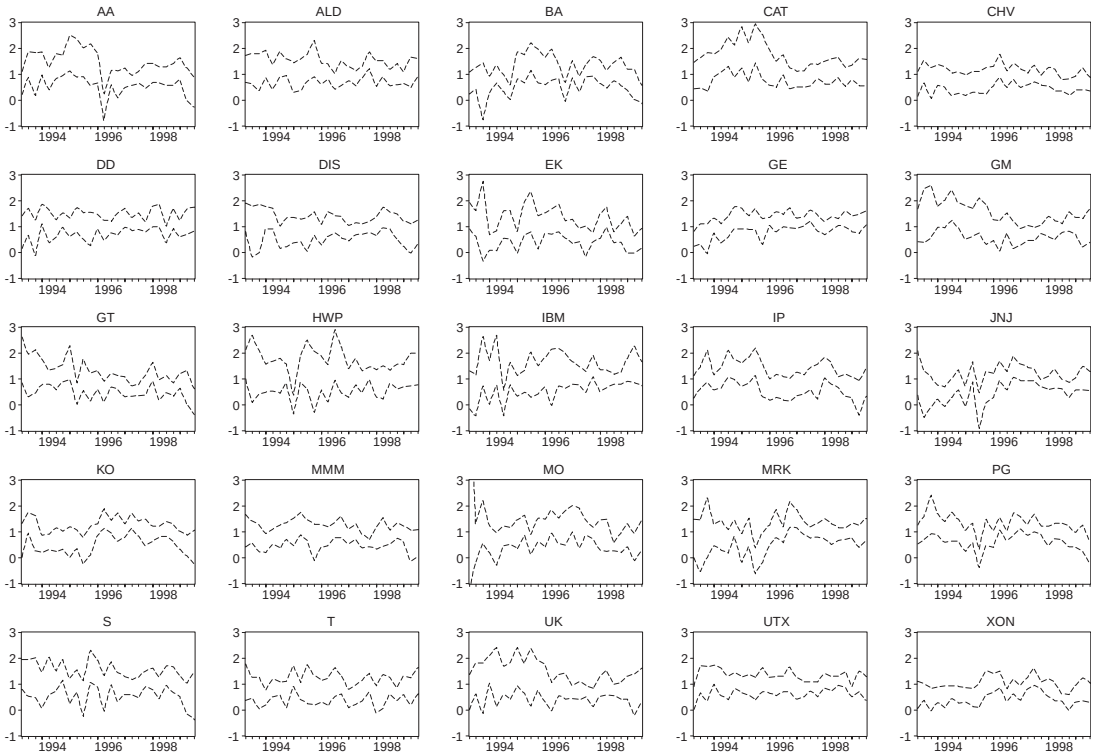


Fig. 14. Ninety-Five Percent Confidence Intervals for Quarterly Beta, Short Sample, Daily Sampling. *Note:* The Time Series of 95 Percent Confidence Intervals for the Underlying Quarterly Integrated Beta, Calculated Using the Results of Barndorff-Nielsen and Shephard (2003) are Shown. The Sample Covers the Period from 1993:2 through 1999:3. We Calculate the Realized Quarterly Betas from Daily Returns.

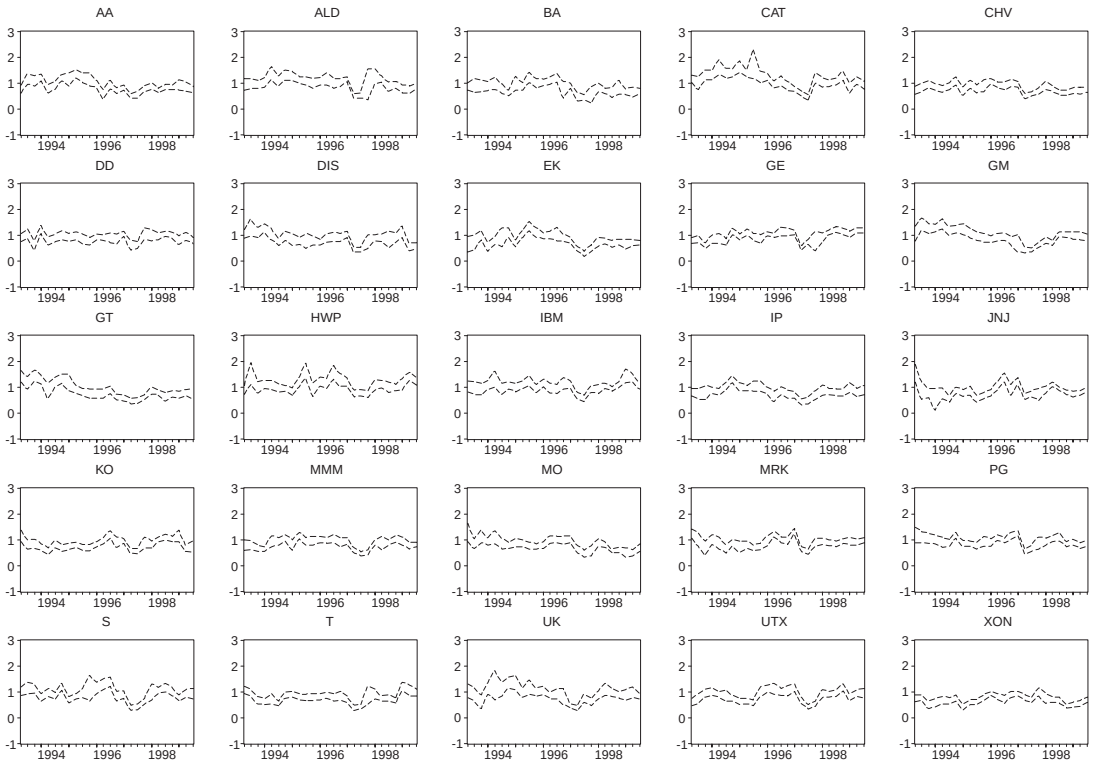


Fig. 15. Ninety-Five Percent Confidence Intervals for Quarterly Beta, Short Sample, 15-Min Sampling. *Note:* The Time Series of 95 Percent Confidence Intervals for the Underlying Quarterly Integrated Beta, Calculated Using the Results of Barndorff-Nielsen and Shephard (2003) are Shown. The Sample Covers the Period from 1993:2 through 1999:3. We Calculate the Realized Quarterly Betas from 15-min Returns.

other words, the high-frequency beta measures importantly complement the results for the betas obtained from the lower frequency daily data, by more clearly highlighting the dynamic evolution of individual security betas. In the web appendix to this paper (www.ssc.upenn.edu/∼fdiebold), we perform a preliminary analysis of realized betas computed from high-frequency data over the shorter 7-year sample period. The results are generally supportive of the findings reported here, but the relatively short sample available for the analysis invariably limits the power of our tests for fractional integration and nonlinear cointegration. In the concluding remarks to this paper, we also sketch a new and powerful econometric framework that we plan to pursue in future, much more extensive, work using underlying high-frequency data.

4.2. HAC Asymptotic Standard Errors

As noted previously, the quarterly realized betas are just regression coefficients computed quarter-by-quarter from CAPM regressions using intra-quarter daily data. One could obtain consistent estimates of the standard errors of those quarterly regression-based betas using HAC approaches, such as Newey–West, under the *very stringent* auxiliary assumption that the period-by-period betas are constant. For comparison to the continuous-record asymptotic bands discussed above, we also compute these HAC standard error bands.

In Fig. 16, we provide the Newey–West 95 percent confidence intervals for the quarterly realized betas. Comparing the figure to Fig. 13, there is not much difference in the assessment of the estimation uncertainly inherent in the quarterly beta measures obtained from the two alternative procedures based on daily data. However, as noted above, there are likely important gains to be had from moving to high-frequency intraday data.

5. SUMMARY, CONCLUDING REMARKS, AND DIRECTIONS FOR FUTURE RESEARCH

We have assessed the dynamics and predictability in realized betas, relative to the dynamics in the underlying market variance and covariances with the market. Key virtues of the approach include the fact that it does not require an assumed volatility model, and that it does not require an assumed model of time variation in beta. We find that, although the realized variances and covariances fluctuate widely and are highly persistent and predictable (as is

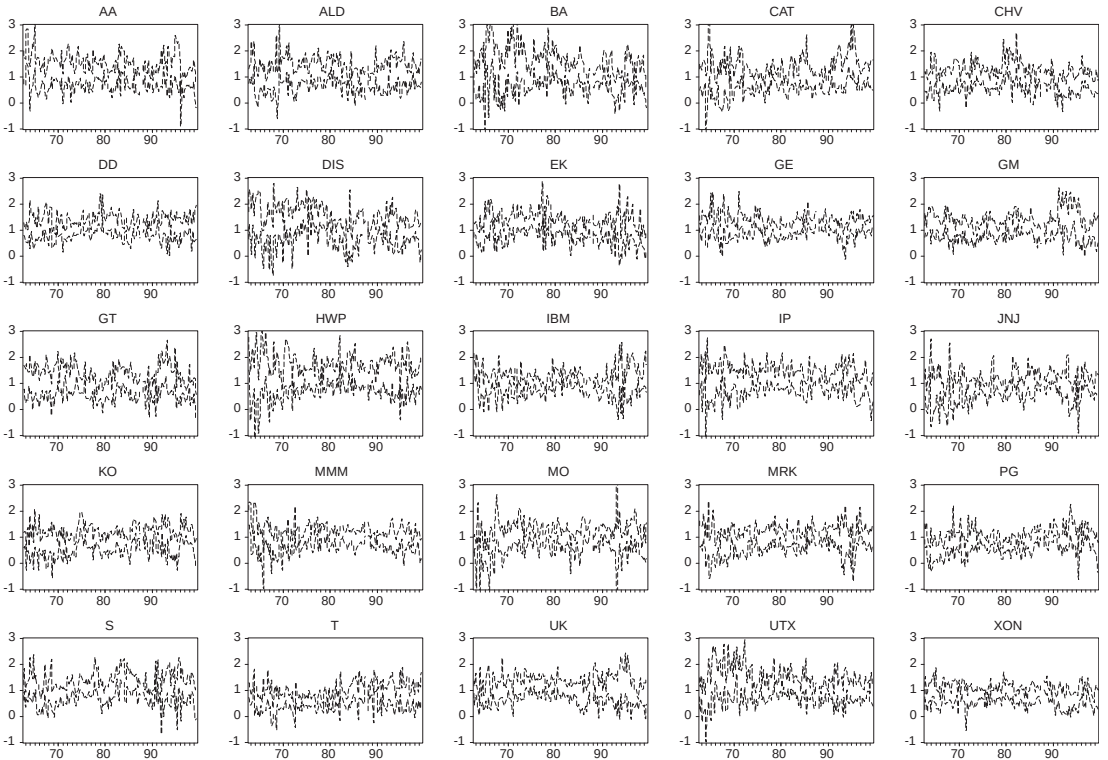


Fig. 16. Ninety-Five Percent Confidence Intervals for Quarterly Beta, Long Sample, Daily Sampling (Newey–West). *Note:* The Time Series of Newey–West 95 Percent Confidence Intervals for the Underlying Quarterly Integrated Beta. The Sample Covers the Period from 1962:3 through 1999:3, with the 1987:4 Outlier Excluded are Shown. We Calculate the Realized Quarterly Betas from Daily Returns.

well-known), the realized betas, which are simple non-linear functions of the realized variances and covariances, display much less persistence and predictability.

The empirical literature on systematic risk measures, as captured by beta, is much too large to be discussed in a sensible fashion here. Before closing, however, we do want to relate our approach and results to the literature on latent factor models and two key earlier papers that have important implications for the potential time variation of betas and the further use of the techniques developed here.

First, our results are closely linked to the literature on the latent factor volatility model, as studied by a number of authors, including Diebold and Nerlove (1989), Harvey, Ruiz, and Shephard (1994), King, Sentana, and Wadhwani (1994), Fiorentini, Sentana, and Shephard (1998), and Jacquier and Marcus (2000). Specifically, consider the model,

$$\boldsymbol{r}_{it} = \beta_i f_t + v_{it}, \quad f_t | I_t \sim (0, h_t) \tag{15a}$$

$$v_{it} \overset{iid}{\sim} (0, \omega_i^2), \quad cov(v_{it} v_{jt'}) = 0, \ \forall i \neq j, \ t \neq t' \tag{15b}$$

where $i, j = 1, \ldots, N$, and $t = 1, \ldots, T$. The ith and jth time-t conditional (on h_t) variances, and the ijth conditional covariance, for arbitrary i and j, are then given by

$$h_{it} = \beta_i^2 h_t + \omega_i^2, \quad h_{jt} = \beta_j^2 h_t + \omega_j^2, \ \ cov_{ijt} = \beta_i \beta_j h_t \tag{16}$$

Assume, as is realistic in financial contexts, that all betas are nonnegative, and consider what happens as h_t increases, say: all conditional variances increase, and all pairwise conditional covariances increase. Hence, the market variance increases, and the covariances of individual equities with the market increase. Two observations are immediate: (1) both the market variance and the covariances of individual equities with the market are time-varying, and (2) because the market variance moves together with the covariances of individual equities with the market, the market betas may not vary as much – indeed in the simple one-factor case sketched here, the betas are constant, by construction! The upshot is that wide fluctuations in the market variance and individual equity covariances with the market, yet no variation in betas, is precisely what one expects to see in a latent (single) factor volatility model. It is also, of course, quite similar to what we found in the data: wide variation and persistence in market variance and individual equity covariances with the market, yet less variation and persistence in betas. Notice, also the remarkable similarity in the correlograms for the individual realized covariances in Fig. 5. This is another indication of a strong coherence in the dynamic evolution of

the individual covariances, consistent with the presence of one dominant underlying factor.

Second, our results also complement and expand upon those of Braun et al. (1995), who study the discrepancy in the time series behavior of betas relative to the underlying variances and covariances for 12 industry portfolios using bivariate Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) models. They also find variation and persistence in the conditional variances and covariances, and less variation and persistence in betas. Moreover, they find the strong asymmetric relationship between return innovations and future return volatility to be entirely absent in the conditional betas.[17] Hence, at the portfolio level they document similar qualitative behavior between the variances and covariances relative to the betas as we do. However, their analysis is linked directly to a specific parametric representation, it studies industry portfolios, and it never contemplates the hypothesis that the constituent components of beta – variances and covariances – may be of a long memory form. This latter point has, of course, been forcefully argued by numerous subsequent studies. Consequently, our investigation can be seen as a substantive extension of their findings performed in a fully nonparametric fashion.

Third, our results nicely complement and expand upon those of Ghysels (1998), who argues that the constant beta CAPM, as bad as it may be, is nevertheless not as bad as some popular conditional CAPMs. We provide some insight into why allowing for time-varying betas may do more harm than good when estimated from daily data, even if the true underlying betas display significant short memory dynamics: it may not be possible to estimate reliably the persistence or predictability in individual realized betas, so good in-sample fits may be spurious artifacts of data mining.[18] We also establish that there should be a real potential for the use of high-frequency intraday data to resolve this dilemma.

In closing, therefore, let us sketch an interesting framework for future research using high-frequency intraday data, which will hopefully deliver superior estimates of integrated volatilities by directly exploiting insights from the continuous-record asymptotics of Barndorff-Nielsen and Shephard (2003). Consider the simple state-space representation:

$$\hat{\beta}_{i,t} = \beta_{i,t} + u_{i,t} \tag{17a}$$

$$\beta_{i,t} = a_0 + a_1 \beta_{i,t-1} + v_{i,t} \tag{17b}$$

$$u_{i,t} \sim N\left(0, \left(\sum_{j=1}^{N_t} \boldsymbol{r}_{mjt}^2\right)^{-2} \hat{g}_{it}\right), \quad v_{i,t} \sim N(0, \sigma_{v,i,t}^2) \tag{17c}$$

The measurement Equation (17a) links the observed realized beta to the unobserved true underlying integrated beta by explicitly introducing a normally distributed error with the asymptotically valid variance obtained from the continuous-record distribution of Barndorff-Nielsen and Shephard (2003). The transition Equation (17b) is a standard first-order autoregression with potentially time-varying error variance.[19] The simplest approach would be to let $v_{i,t}$ have a constant variance, but it is also straightforward to let the variance change with the underlying variability in the realized beta measure, so that the beta innovations become more volatile as the constituent parts, the market variance and the covariance of the stock return with the market, increase. This approach directly utilizes the advantages of high-frequency intraday beta measurements by incorporating estimates of the measurement errors to alleviate the errors-n-variables problem, while explicitly recognizing the heteroskedasticity in the realized beta series. We look forward to future research along these lines.

NOTES

1. The Roll (1977) critique is also relevant. That is, even if we somehow knew what factor(s) should be priced, it is not clear that the factor proxies measured in practice would correspond to the factor required by the theory.

2. See Keim and Hawawini (1999) for a good discussion of the difficulty of interpreting additional empirically motivated factors in terms of systematic risk.

3. There are of course qualifications, notably Ghysels (1998), which we discuss subsequently.

4. The idea of conditioning in the CAPM is of course not unrelated to the idea of multi-factor pricing mentioned earlier.

5. The underlying theory and related empirical strategies are developed in Andersen, Bollerslev, Diebold, and Labys (2001b, 2003), Andersen, Bollerslev, Diebold, and Ebens (2001a), Andersen and Bollerslev (1998), and Barndorff-Nielsen and Shephard (2003). Here, we sketch only the basics; for a more rigorous treatment in the framework of special semimartingales, see the survey and unification by Andersen et al. (2005).

6. This notion of integrated volatility already plays a central role in the stochastic volatility option pricing literature, in which the price of an option typically depends on the distribution of the integrated volatility process for the underlying asset over the life of the option. See, for example, the well-known contribution of Hull and White (1987).

7. Formal theoretical asymptotic justification for this finding has very recently been provided by Barndorff-Nielsen and Shephard (2004).

8. One could, of course, attempt a linear cointegration approach by taking logs of the realized volatilities and covariances, but there is no theoretical reason to expect all covariances to be positive, and our realized covariance measures are indeed sometimes negative, making logarithmic transformations problematic.

9. We compute the quarterly realized variance, covariances, and betas from slightly different numbers of observations due to the different numbers of trading days across the quarters.

10. Note also that the Dickey–Fuller statistics indicate that unit roots are not present in the market variance, individual equity covariances with the market, or market betas, despite their persistent dynamics.

11. For the realized covariances and realized betas, we show the median autocorrelations functions.

12. The standard error band (under the null of an i.i.d. series) indicated in Fig. 4 is only valid for the realized market variance. It should be lower for the two other series, reflecting the effective averaging in constructing the median values. In fact, it should be considerably lower for the beta series due to the near uncorrelated nature of the underlying beta dynamics, while the appropriate reduction for the covariance series would be less because of the strong correlation across the series. We cannot be more precise on this point without imposing some direct assumptions on the correlation structure across the individual series.

13. A partial list of references not written by the present authors includes Breidt, Crato, and de Lima (1998), Comte and Renault (1998), Harvey (1998), and Robinson (2001), as well as many of the earlier papers cited in Baillie (1996).

14. Note that only one figure is needed, despite the many different realized covariances, because all $G(j)$ are identical, as all processes are assumed to be $ARFIMA(0, 0.42, 0)$.

15. N has a time subscript because the number of trading days varies slightly across quarters.

16. The high-frequency tick-by-tick data underlying the 15-min returns was obtained from the TAQ (Trade And Quotation) database. We refer the reader to the web appendix to this paper, available at www.ssc.upenn.edu/~fdiebold, for a more detailed description of the data capture, return construction, and high-frequency beta measurements.

17. In contrast, on estimating a similar EGARCH model for individual daily stock returns, Cho and Engle (2000) find that daily company specific betas *do* respond asymmetrically to good and bad news.

18. Chang and Weiss (1991) argue that a strictly stationary Autoregressive Moving Average (ARMA) (1,1) model provides an adequate representation for most individual quarterly beta series. Their sample size is smaller than ours, however, and their estimated betas are assumed to be constant over each quarter. Also, they do not provide any separate consideration of the persistence of the market variance or the individual covariances with the market.

19. Generalization to an arbitrary ARMA process or other stationary structures for the evolution in the true betas is, of course, straightforward.

ACKNOWLEDGMENTS

For useful discussion we thank participants at the UCSD Conference on Predictive Methodology and Applications in Economics and Finance (in honor of Granger), and the London School of Economics Fifth Financial Markets Group Conference on Empirical Finance. We also thank seminar participants at the University of Pennsylvania, as well as Andrew Ang, Michael Brandt, Mike Chernov, Graham Elliott, Eric Ghysels, Rich Lyons, Norman Swanson, and Mark Watson.

REFERENCES

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*, 885–905.

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2005). Parametric and nonparametric volatility measurement. In: L. P. Hansen & Y. Ait-Sahalia (Eds), *Handbook of financial econometrics*. Amsterdam: North-Holland.

Andersen, T., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001a). The distribution of realized stock return volatility. *Journal of Financial Economics*, *61*, 43–76.

Andersen, T., Bollerslev, T., Diebold, F. X., & Labys, P. (2001b). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, *96*, 42–55.

Andersen, T., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*, 579–626.

Andersen, T., Bollerslev, T., & Meddahi, N. (2004). Analytic evaluation of volatility forecasts. *International Economic Review*, *45*, 1079–1110.

Andrews, D. W. K., & Guggenberger, P. (2003). A bias-reduced log-periodogram regression estimator of the long-memory parameter. *Econometrica*, *71*, 675–712.

Ang, A., & Chen, J. C. (2003). *CAPM over the long run: 1926–2001*. Manuscript, Columbia University and University of Southern California.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, *73*, 5–59.

Barndorff-Nielsen, O. E., & Shephard, N. (2003). *Econometric analysis of realized covariation: High frequency covariance, regression and correlation in financial economics*. Manuscript, Oxford: Nuffield College.

Barndorff-Nielsen, O. E., & Shephard, N. (2004). *A feasible central limit theory for realised volatility under leverage*. Manuscript, Oxford: Nuffield College.

Bollerslev, T., Engle, R. F., & Wooldridge, J. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, *96*, 113–131.

Braun, P. A., Nelson, D. B., & Sunier, A. M. (1995). Good news, bad news, volatility, and betas. *Journal of Finance*, *50*, 1575–1603.

Breidt, F. J., Crato, N., & de Lima, P. (1998). On the detection and estimation of long memory in stochastic volatility. *Journal of Econometrics*, *73*, 325–334.

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Campbell, J. Y., & Vuolteenaho, T. (2004). Bad beta, good beta. *American Economic Review*, *94*, 1249–1275.

Chang, W.-C., & Weiss, D. E. (1991). An examination of the time series properties of beta in the market model. *Journal of the American Statistical Association*, *86*, 883–890.

Cheung, Y.-W., & Lai, K. S. (1993). A fractional cointegration analysis of purchasing power parity. *Journal of Business and Economic Statistics*, *11*, 103–112.

Cho, Y. H., & Engle, R. F. (2000). *Time-varying and asymmetric effects of news: Empirical analysis of blue chip stocks*. Manuscript, Rutgers University and New York University.

Cohen, R., Polk, C., & Vuolteenaho, T. (2002). *Does risk or mispricing explain the cross-section of stock prices*. Manuscript, Kellogg School, Northwestern University.

Comte, F., & Renault, E. (1998). Long memory in continuous time stochastic volatility models. *Mathematical Finance*, *8*, 291–323.

Diebold, F. X., & Kilian, L. (2001). Measuring predictability: Theory and macroeconomic implications. *Journal of Applied Econometrics*, *16*, 657–669.

Diebold, F. X., & Nerlove, M. (1989). The dynamics of exchange rate volatility: A multivariate latent factor ARCH model. *Journal of Applied Econometrics*, *4*, 1–22.

Dybvig, P. H., & Ross, S. A. (1985). Differential information and performance measurement using a security market line. *Journal of Finance*, *40*, 383–400.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, *55*, 251–276.

Engle, R. F., & Kozicki, S. (1993). Testing for common features. *Journal of Business and Economic Statistics*, *11*, 369–383.

Fama, E. F. (1976). *Foundations of finance*. New York: Basic Books.

Fama, E. F., & French, K. R. (1992). The cross section of expected stock returns. *Journal of Finance*, *47*, 427–465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3–56.

Ferson, W. E., & Harvey, C. R. (1991). The variation of economic risk premiums. *Journal of Political Economy*, *99*, 385–415.

Ferson, W. E., Kandel, S., & Stambaugh, R. F. (1987). Tests of asset pricing with time-varying expected risk premiums and market betas. *Journal of Finance*, *42*, 201–220.

Fiorentini, G., Sentana, E., & Shephard, N. (1998). *Exact likelihood-based estimation of conditionally heteroskedastic factor models*. Working Paper. University of Alcante, CEMFI and Oxford University.

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, *4*, 221–238.

Ghysels, E. (1998). On stable factor structures in the pricing of risk: Do time-varying betas help or hurt?. *Journal of Finance*, *53*, 549–573.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, *16*, 121–130.

Granger, C. W. J. (1995). Modeling nonlinear relationships between extended-memory variables. *Econometrica*, *63*, 265–279.

Granger, C. W. J., & Newbold, P. (1986). *Forecasting economic time series* (2nd edition). Orlando: Academic Press.

Hansen, L. P., & Richard, S. F. (1987). The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica*, *55*, 587–613.

Harvey, A. C. (1998). Long memory in stochastic volatility. In: J. Knight & S. Satchell (Eds), *Forecasting volatility in financial markets*. London, UK: Butterworth-Heinemann.

Harvey, A. C., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, *61*, 247–264.

Horvath, M. T. K., & Watson, M. W. (1995). Testing for cointegration when some of the cointegrating vectors are prespecified. *Econometric Theory*, *11*, 952–984.

Huang, C., & Litzenberger, R. H. (1988). *Foundations for financial economics*. Amsterdam: North-Holland.

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance*, *42*, 381–400.

Jacquier, E., & Marcus, A. J. (2000). *Market volatility and asset correlation structure*. Working Paper, Boston College.

Jagannathan, R., & Wang, Z. (1996). The conditional CAPM and the cross section of expected returns. *Journal of Finance*, *51*, 3–53.

Keim, D., & Hawawini, G. (1999). *The cross section of common stock returns: A review of the evidence and some new findings*. Manuscript, Wharton School.

King, M., Sentana, E., & Wadhwani, S. (1994). Volatility and links between national stock markets. *Econometrica*, *62*, 901–933.

Lintner, J. (1965a). Security prices, risk and maximal gains from diversification. *Journal of Finance*, *20*, 587–615.

Lintner, J. (1965b). The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, *47*, 13–37.

Meddahi, N. (2002). A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics*, *17*, 479–508.

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

Robinson, P. M., & Marinucci, D. (2001). Semiparametric fractional cointegration analysis. *Journal of Econometrics*, *105*, 225–247.

Roll, R. (1977). A critique of the asset pricing theory's tests. *Journal of Financial Economics*, *4*, 129–176.

Rosenberg, B. (1973). Random coefficient models: The analysis of a cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement*, *2*, 399–428.

Schaefer, S., Brealey, R., Hodges, S., & Thomas, H. (1975). Alternative models of systematic risk. In: E. Elton & M. Gruber (Eds), *International capital markets: An inter- and intra-country analysis* (pp. 150–161). Amsterdam: North-Holland.

Schwert, G. W. (1989). Why does stock market volatility change over time?. *Journal of Finance*, *44*, 1115–1153.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science*, *9*, 227–293.

Wang, K. Q. (2003). Asset pricing with conditioning information: A new test. *Journal of Finance*, *58*, 161–196.

This page intentionally left blank

ASYMMETRIC PREDICTIVE ABILITIES OF NONLINEAR MODELS FOR STOCK RETURNS: EVIDENCE FROM DENSITY FORECAST COMPARISON

Yong Bao and Tae-Hwy Lee

ABSTRACT

We investigate predictive abilities of nonlinear models for stock returns when density forecasts are evaluated and compared instead of the conditional mean point forecasts. The aim of this paper is to show whether the in-sample evidence of strong nonlinearity in mean may be exploited for out-of-sample prediction and whether a nonlinear model may beat the martingale model in out-of-sample prediction. We use the Kullback–Leibler Information Criterion (KLIC) divergence measure to characterize the extent of misspecification of a forecast model. The reality check test of White (2000) using the KLIC as a loss function is conducted to compare the out-of-sample performance of competing conditional mean models. In this framework, the KLIC measures not only model specification error but also parameter estimation error, and thus we treat both types of errors as loss. The conditional mean models we use for the daily closing S&P 500 index returns include the martingale difference,

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 41–62
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20021-X

ARMA, STAR, SETAR, artificial neural network, and polynomial models. Our empirical findings suggest the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric in the sense that the right tails of the return series are predictable via many of the nonlinear models, while we find no such evidence for the left tails or the entire distribution.

1. INTRODUCTION

While there has been some evidence that financial returns may be predictable (see, e.g., Lo & MacKinlay, 1988; Wright, 2000), it is generally believed that financial returns are very close to a martingale difference sequence (MDS). The evidence against MDS is usually stronger from in-sample specification tests than from out-of-sample predictability tests using standard evaluation criteria such as the mean squared forecast error (MSFE) and mean absolute forecast error (MAFE).

In this paper, we investigate if this remains true when we evaluate forecasting models in terms of density forecasts instead of the conditional mean point forecasts using MSFE and MAFE. We examine if the evidence and its significance of the nonlinear predictability of financial returns depend on whether we use point forecast evaluation criteria (MSFE and MAFE) or we use the probability density forecasts. As Clements and Smith (2000, 2001) show, traditional measures such as MSFE may mask the superiority of nonlinear models, whose predictive abilities may be more evident through density forecast evaluation.

Motivated by the encouraging results of Clements and Smith (2000, 2001), we compare the density forecasts of various linear and nonlinear models for the conditional mean of the S&P 500 returns by using the method of Bao, Lee, and Saltoglu (2004, BLS henceforth), where the Kullback and Leibler's (1951) Information Criterion (KLIC) divergence measure is used for characterizing the extent of misspecification of a density forecast model. In BLS's framework, the KLIC captures not only model specification error but also parameter estimation error. To compare the performance of density forecast models in the tails of stock return distributions, we also follow BLS by using the censored likelihood functions to compute the tail minimum KLIC. The reality check test of White (2000) is then constructed using the KLIC as a loss function. We find that, for the entire distribution and the left tails, the S&P 500 daily closing returns are not predictable via various linear

or nonlinear models and the MDS model performs best for out-of-sample forecasting. However, from the right tail density forecast comparison of the S&P 500 data, we find, surprisingly, that the MDS model is dominated by many nonlinear models. This suggests that the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric.

This paper proceeds as follows. In Section 2, we examine the nature of the in-sample nonlinearity using the generalized spectral test of Hong (1999). Section 3 presents various linear and nonlinear models we use for the out-of-sample analysis. In Section 4, we compare these models for the S&P 500 return series employing the density forecast approach of BLS. Section 5 concludes. Throughout, we define $y_t = 100(\ln P_t - \ln P_{t-1})$, where P_t is the S&P 500 index at time t.

2. IN-SAMPLE TEST FOR MARTINGALE DIFFERENCE

We will first explore serial dependence (i.e., any departure from IID) in the S&P 500 returns using Hong's (1999) generalized spectrum. In particular, we are interested in finding significant and predictable nonlinearity in the conditional mean even when the returns are linearly unpredictable.

The basic idea is to transform a strictly stationary series y_t to $e^{\mathbf{i}uy_t}$ and consider the covariance function between the transformed variables $e^{\mathbf{i}uy_t}$ and $e^{\mathbf{i}vy_{t-|j|}}$

$$\sigma_j(u,v) \equiv \mathrm{cov}(e^{\mathbf{i}uy_t}, e^{\mathbf{i}vy_{t-|j|}}) \tag{1}$$

where $\mathbf{i} \equiv \sqrt{-1}$, $u, v \in (-\infty, \infty)$, and $j = 0, \pm 1, \ldots$. Suppose that $\{y_t\}_{t=1}^T$ has a marginal characteristic function $\varphi(u) \equiv \mathbb{E}(e^{\mathbf{i}uy_t})$ and a pairwise joint characteristic function $\varphi_j(u,v) \equiv \mathbb{E}(e^{\mathbf{i}(uy_t+vy_{t-|j|})})$. Straightforward algebra yields $\sigma_j(u,v) = \varphi_j(u,v) - \varphi(u)\varphi(v)$. Because $\varphi_j(u,v) = \varphi(u)\varphi(v)$ for all u, v if and only if y_t and $y_{t-|j|}$ are independent, $\sigma_j(u, v)$ can capture any type of pairwise serial dependence over various lags.

When $\sup_{u,v\in(-\infty,\infty)} \sum_{j=-\infty}^{\infty} |\sigma_j(u,v)| < \infty$, the Fourier transform of $\sigma_j(u, v)$ exists

$$f(\omega,u,v) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j(u,v) e^{-\mathbf{i}j\omega}, \quad \omega \in [-\pi, \pi] \tag{2}$$

Like $\sigma_j(u, v)$, $f(\omega, u, v)$ can capture all pairwise serial dependencies in $\{y_t\}$ over various lags. Hong (1999) calls $f(\omega, u, v)$ a "generalized spectral

density'' of $\{y_t\}$, and shows that $f(\omega, u, v)$ can be consistently estimated by

$$\hat{f}_n(\omega, u, v) \equiv \frac{1}{2\pi} \sum_{j=1-n}^{n-1} (1 - |j|/n)^{1/2} k(j/p) \hat{\sigma}_j(u, v) e^{-\mathbf{i}j\omega} \tag{3}$$

where $\hat{\sigma}_j(u, v) \equiv \hat{\varphi}_j(u, v) - \hat{\varphi}_j(u, 0)\hat{\varphi}_j(0, v)$ is the empirical generalized covariance, $\hat{\varphi}_j(u, v) \equiv (n - |j|)^{-1} \sum_{t=|j|+1}^{n} e^{\mathbf{i}(uy_{t+v}y_{t-|j|})}$ is the empirical pairwise characteristic function, $p \equiv p_n$ a bandwidth or lag order, and $k(\cdot)$ a kernel function or ''lag window''. Commonly used kernels include the Bartlett, Daniell, Parzen, and Quadratic–Spectral kernels.

When $\{y_t\}$ is IID, $f(\omega, u, v)$ becomes a ''flat'' generalized spectrum:

$$f_0(\omega, u, v) \equiv \frac{1}{2\pi} \sigma_0(u, v), \quad \omega \in [-\pi, \pi]$$

Any deviation of $f(\omega, u, v)$ from the flat spectrum $f_0(\omega, u, v)$ is the evidence of serial dependence. Thus, to detect serial dependence, we can compare $\hat{f}_n(\omega, u, v)$ with the estimator

$$\hat{f}_0(\omega, u, v) \equiv \frac{1}{2\pi} \hat{\sigma}_0(u, v), \quad \omega \in [-\pi, \pi]$$

To explore the nature of serial dependence, one can compare the derivative estimators

$$\hat{f}_n^{(0,m,l)}(\omega, u, v) \equiv \tfrac{1}{2\pi} \sum_{j=1-n}^{n-1} (1 - |j|/n)^{1/2} k(j/p) \hat{\sigma}_j^{(m,l)}(u, v) e^{-\mathbf{i}j\omega}$$

$$\hat{f}_0^{(0,m,l)}(\omega, u, v) \equiv \tfrac{1}{2\pi} \hat{\sigma}_0^{(m,l)}(u, v)$$

where $\hat{\sigma}_j^{(m,l)}(u, v) \equiv \partial^{m+l} \hat{\sigma}_j(u, v) / \partial^m u \partial^l v$ for $m, l \geq 0$. Just as the characteristic function can be differentiated to generate various moments, generalized spectral derivatives can capture various specific aspects of serial dependence, thus providing information on possible types of serial dependence.

Hong (1999) proposes a class of tests based on the quadratic norm

$$\begin{aligned} Q\left(\hat{f}_n^{(0,m,l)}, \hat{f}_0^{(0,m,l)}\right) &\equiv \int \int_{-\pi}^{\pi} \left| \hat{f}_n^{(0,m,l)}(\omega, u, v) - \hat{f}_0^{(0,m,l)}(\omega, u, v) \right|^2 \mathrm{d}\omega \, \mathrm{d}W_1(u) \, \mathrm{d}W_2(v) \\ &= \frac{2}{\pi} \int \sum_{j=1}^{n-1} k^2(j/p)(1 - j/n) \left| \hat{\sigma}_j^{(m,l)}(u, v) \right|^2 \mathrm{d}W_1(u) \, \mathrm{d}W_2(v) \end{aligned}$$

where the second equality follows by Parseval's identity, and the unspecified integrals are taken over the support of $W_1(\cdot)$ and $W_2(\cdot)$, which are positive

and nondecreasing weighting functions that set weight about zero equally. The generalized spectral test statistic $M(m, l)$ is a standardized version of the quadratic norm. Given (m, l), $M(m, l)$ is asymptotically one-sided $N(0,1)$ under the null hypothesis of serial independence, and thus the upper-tailed asymptotic critical values are 1.65 and 2.33 at the 5% and 1% levels, respectively.

We may first choose $(m, l) = (0, 0)$ to check if there exists any type of serial dependence. Once generic serial dependence is discovered using $M(0,0)$, we may use various combinations of (m, l) to check specific types of serial dependence. For example, we can set $(m, l) = (1, 0)$ to check whether there exists serial dependence in mean. This checks whether $\mathbb{E}(y_t|y_{t-j}) = \mathbb{E}(y_t)$ for all $j > 0$, and so it is a suitable test for the MDS hypothesis. It can detect a wide range of deviations from MDS. To explore whether there exists linear dependence in mean, we can set $(m, l) = (1, 1)$. If $M(1,0)$ is significant but $M(1,1)$ is not, we can speculate that there may exist only nonlinear dependence in mean. We can go further to choose $(m, l) = (1, l)$ for $l = 2, 3, 4$, testing if $\text{cov}(y_t, y_{t-j}^l) = 0$ for all $j > 0$. These essentially check whether there exist ARCH-in-mean, skewness-in-mean, and kurtosis-in-mean effects, which may arise from the existence of time-varying risk premium, asymmetry, and improper account of the concern over large losses, respectively. Table 1 lists a variety of spectral derivative tests and the types of dependence they can detect, together with the estimated $M(m, l)$ statistics.[1]

We now use the generalized spectral test to explore serial dependence of the daily S&P 500 closing return series, retrieved from *finance.yahoo.com*. They are from January 3, 1990 to June 30, 2003 ($T = 3403$).

The statistic $M(m, l)$ involves the choice of a bandwidth p in its computation, see Hong (1999, p. 1204). Hong proposes a data-driven method to choose p. This method still involves the choice of a preliminary bandwidth $\bar{p}$. Simulations in Hong (1999) show that the choice of $\bar{p}$ is less important than that of p. We consider $\bar{p}$ in the range 6–15 to examine the robustness of $M(m, l)$ with respect to the choice of $\bar{p}$. We use the Daniell kernel, which maximizes the asymptotic power of $M(m, l)$ over a class of kernels. We have also used the Bartlett, Parzen, and Quadratic–Spectral kernels, whose results are similar to those based on the Daniell kernel and are not reported in this paper.

Table 1 reports the values of $M(m, l)$ for $\bar{p} = 6, 9, 12, 15$. The results for various values of $\bar{p}$ are quite similar. $M(m, l)$ has an asymptotic one-sided $N(0,1)$ distribution, so the asymptotic critical value at the 5% level is 1.65. The $M(0,0)$ statistic suggests that the random walk hypothesis is strongly rejected. In contrast, the correlation test $M(1,1)$ is insignificant, implying

Table 1. Generalized Spectral Tests.

Test	Statistic $M(m, l)$	Test Function $\sigma_j^{(m,l)}(u,v)$	Preliminary Bandwidth $\bar{p}=6$	$\bar{p}=9$	$\bar{p}=12$	$\bar{p}=15$
IID	$M(0, 0)$	$\sigma_j = (u, v)$	51.02	58.23	63.75	67.85
MDS	$M(1, 0)$	$\text{cov}(y_t, e^{\mathbf{i}vy_{t-j}})$	17.40	18.28	18.67	19.04
Correlation	$M(1, 1)$	$\text{cov}(y_t, y_{t-j})$	−0.10	0.44	0.63	0.68
ARCH-in-mean	$M(1, 2)$	$\text{cov}(y_t, y_{t-j}^2)$	56.24	55.83	55.36	54.84
Skewness-in-mean	$M(1, 3)$	$\text{cov}(y_t, y_{t-j}^3)$	−0.11	−0.38	−0.50	−0.51
Kurtosis-in-mean	$M(1, 4)$	$\text{cov}(y_t, y_{t-j}^4)$	29.85	29.99	29.57	29.18
Nonlinear ARCH	$M(2, 0)$	$\text{cov}(y_t^2, e^{\mathbf{i}vy_{t-j}})$	62.15	70.75	76.71	81.30
Leverage	$M(2, 1)$	$\text{cov}(y_t^2, y_{t-j})$	9.25	8.57	8.52	8.57
Linear ARCH	$M(2, 2)$	$\text{cov}(y_t^2, y_{t-j}^2)$	172.87	182.41	188.53	193.64
Conditional skewness	$M(3, 0)$	$\text{cov}(y_t^3, e^{\mathbf{i}vy_{t-j}})$	7.63	6.98	6.64	6.36
Conditional skewness	$M(3, 3)$	$\text{cov}(y_t^3, y_{t-j}^3)$	27.82	26.66	26.69	26.83
Conditional kurtosis	$M(4, 0)$	$\text{cov}(y_t^4, e^{\mathbf{i}vy_{t-j}})$	17.16	18.17	19.12	20.10
Conditional kurtosis	$M(4, 4)$	$\text{cov}(y_t^4, y_{t-j}^4)$	35.56	35.22	35.22	35.25

Note: All generalized spectral test statistics $M(m, l)$ are asymptotically one-sided $N(0, 1)$ and thus upper-tailed asymptotic critical values are 1.65 and 2.33 at the 5% and 1% levels, respectively. $M(0, 0)$ is to check if there exists any type of serial dependence. $M(1, 0)$ is to check whether there exists serial dependence in mean. To explore whether there exists linear dependence in mean, we can set $(m, l) = (1, 1)$. If $M(1, 0)$ is significant but $M(1, 1)$ is not, we can speculate that there may exist only nonlinear dependence in mean. We choose $(m, l) = (1, l)$ with $l = 2, 3, 4$, to test if $\mathbb{E}(y_t|y_{t-j}^l) = 0$ for all $j>0$. The PN model is to exploit the nonlinear predictive evidence of polynomials found from $M(1, l)$ with $l = 2, 3, 4$.

that $\{y_t\}$ is an uncorrelated white noise. This, however, does not necessarily imply that $\{y_t\}$ is an MDS. Indeed, the martingale test $M(1,0)$ strongly rejects the martingale hypothesis as its statistic is above 17. This implies that the S&P 500 returns, though serially uncorrelated, has a nonzero mean conditional on its past history. Thus, suitable nonlinear time series models may be able to predict the future returns. The polynomial (PN) model (to be discussed in the next section) is to exploit the nonlinear predictive evidence of the lth power of returns, as indicated by the $M(1,l)$ statistics.

The test $M(2,0)$ shows possibly nonlinear time-varying volatility, and the linear ARCH test $M(2,2)$ indicates very strong linear ARCH effects. We also observe that the leverage effect ($M(2,1)$) is significant and there exist significant conditional skewness as evidenced from $M(3,0)$ and $M(3,3)$, and large conditional kurtosis as evidenced from $M(4,0)$ and $M(4,4)$.

It is important to explore the implications of these in-sample findings of nonlinearity in the conditional mean. The fact that the S&P 500 return series

is not an MDS implies it may be predictable in the conditional mean. In the next section, we will use various linear and nonlinear time series models to examine this issue.

3. CONDITIONAL MEAN MODELS

Let $\mathscr{F}_{t-1}$ be the information set containing information about the process $\{y_t\}$ up to and including time $t-1$. Since our interest is to investigate the predictability of stock returns in the conditional mean $\mu_t = \mathbb{E}(y_t|\mathscr{F}_{t-1})$, we assume that y_t is conditionally normally distributed and the conditional variance $\sigma_t^2 = \mathbb{E}(\varepsilon_t^2|\mathscr{F}_{t-1})$ follows a GARCH(1,1) process $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$, where $\varepsilon_t = y_t - \mu_t$. We consider the following nine models for μ_t in three classes:

(i) the MDS model

$$\text{MDS} \qquad y_t = \varepsilon_t$$

(ii) four linear autoregressive moving average (ARMA) models

Constant	$y_t = a_0 + \varepsilon_t$
MA(1)	$y_t = a_0 + b_1\varepsilon_{t-1} + \varepsilon_t$
ARMA(1, 1)	$y_t = a_0 + a_1 y_{t-1} + b_1\varepsilon_{t-1} + \varepsilon_t$
AR(2)	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$

(iii) four nonlinear models, namely, the polynomial (PN), neural network (NN), self-exciting transition autoregressive (SETAR), and smooth transition autoregressive (STAR) models,

PN(2, 4)	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \sum_{j=1}^{2}\sum_{i=2}^{4} a_{ij} y_{t-j}^i + \varepsilon_t$
NN(2, 5)	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \sum_{i=1}^{5} \delta_i G(\gamma_{0i} + \sum_{j=1}^{2} \gamma_{ji} y_{t-j}) + \varepsilon_t$
SETAR	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + (b_0 + b_1 y_{t-1} + b_2 y_{t-2})\mathbf{1}(y_{t-1} > c) + \varepsilon_t$
STAR	$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + (b_0 + b_1 y_{t-1} + b_2 y_{t-2}) G(\gamma(y_{t-1} - c)) + \varepsilon_t$

where $G(z) = 1/(1 + e^{-z})$ is a logistic function and $\mathbf{1}(\cdot)$ denotes an indicator function that takes 1 if its argument is true and 0 otherwise. Note that the four nonlinear models nest the AR(2) model.

All the above models have been used in the literature, with apparently mixed results on the predictability of stock returns. Hong and Lee (2003) use the AR, PN, and NN models. McMillan (2001) and Kanas (2003) find evidence supporting the NN and STAR models, while Bradley and Jansen (2004) find no evidence for the STAR model. We note that these authors use the MSFE criterion for out-of-sample forecasting evaluation. Racine (2001) finds no predictability evidence using the NN model. Anderson, Benzoni, and Lund (2002) use the MA model for estimation.

The results from the generalized spectral test reported in Table 1 suggest that $\mathbb{E}(y_t|\mathcal{F}_{t-1})$ is time-varying in a nonlinear manner, because $M(1,1)$ is insignificant but $M(1,0)$ is significant. Also, we note that the $M(1,l)$ statistics are significant with $l = 2, 4$ but not with $l = 1, 3$, indicating volatility and tail observations may have some predictive power for the returns but not the skewness of the return distribution. The PN model is to exploit this nonlinear predictive evidence of the lth order power of the lagged returns.

4. OUT-OF-SAMPLE TEST FOR MARTINGALE DIFFERENCE

We now examine if the in-sample evidence of nonlinear predictability of the S&P 500 returns from the generalized spectral test in the previous section may be carried over to the out-of-sample forecasting. While the in-sample generalized spectral test does not involve parameter estimation of a particular nonlinear model, the out-of-sample test requires the estimation of model parameters since it is based on some particular choice of nonlinear models. Despite the strong nonlinearity found in the conditional mean from the in-sample tests, model uncertainty and parameter uncertainty usually make the out-of-sample results much weaker than the in-sample nonlinear evidence, see Meese and Rogoff (1983).

Given model uncertainty, econometricians tend to search for a proper model over a large set of candidate models. This can easily cause the problem of data snooping, see Lo and MacKinlay (1990). In order to take care of the data snooping bias, we follow the method of White (2000) in our comparison of multiple competing models. Moreover, we use the KLIC

measure as our loss function in the comparison. As emphasized by BLS (2004), the KLIC measure captures both the loss due to model specification error and the loss due to parameter estimation error. It is important to note that in comparing forecasting models we treat parameter estimation error as a loss. Now, we briefly discuss the BLS test.[2]

4.1. The BLS Test

Suppose that $\{y_t\}$ has a true, unknown conditional density function $f(y_t) \equiv f(y_t|\mathscr{F}_{t-1})$. Let $\psi(y_t;\boldsymbol{\theta}) = \psi(y_t|\mathscr{F}_{t-1};\boldsymbol{\theta})$ be a one-step-ahead conditional density forecast model with parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a finite-dimensional vector of parameters in a compact parameter space $\boldsymbol{\Theta}$. If $\psi(\cdot;\boldsymbol{\theta}_0) = f(\cdot)$ for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then the one-step-ahead density forecast is correctly specified and hence optimal, as it dominates all other density forecasts for any loss function (e.g., Diebold, Gunther, & Tay, 1998; Granger & Pesaran, 2000a, b). As our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we consider two subsamples $\{y_t\}_{t=1}^{R}$ and $\{y_t\}_{t=R+1}^{T}$: we use the first sample to estimate the unknown parameter vector $\boldsymbol{\theta}$ and the second sample to compare the out-of-sample density forecasts.

In practice, it is rarely the case that we can find an optimal model as all the models can be possibly misspecified. Our task is then to investigate which model can approximate the true model most closely. We have to first define a metric to measure the distance of a given model to the truth and then compare different models in terms of this distance. The adequacy of a postulated distribution may be appropriately measured by the KLIC divergence measure between $f(\cdot)$ and $\psi(\cdot)$: $I(f:\psi,\boldsymbol{\theta}) = \mathbb{E}\left[\ln f(y_t) - \ln\psi(y_t;\boldsymbol{\theta})\right]$, where the expectation is with respect to the true distribution. Following Vuong (1989), we define the distance between a model and the true density as the minimum of the KLIC

$$I(f:\psi,\boldsymbol{\theta}^*) = \mathbb{E}\left[\ln f(y_t) - \ln\psi(y_t;\boldsymbol{\theta}^*)\right] \tag{4}$$

and $\boldsymbol{\theta}^*$ is the pseudo-true value of $\boldsymbol{\theta}$, the parameter value that gives the minimum $I(f:\psi,\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$ (e.g., White, 1982). The smaller this distance, the closer the model $\psi(\cdot;\boldsymbol{\theta})$ is to the true density. Thus, we can use this measure to compare the relative distance of a battery of competing models to the true model $f_t(\cdot)$. However, $I(f:\psi,\boldsymbol{\theta}^*)$ is generally unknown, since we cannot observe $f(\cdot)$ and thereby we can not evaluate the expectation in (4). Under some regularity conditions, it can nevertheless be shown that

$\mathbb{E}\left[\ln f(y_t) - \ln\psi(y_t;\boldsymbol{\theta}^*)\right]$ can be consistently estimated by

$$\hat{I}(f:\psi) = \frac{1}{n}\sum_{t=R+1}^{T}\left[\ln f(y_t) - \ln\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})\right] \tag{5}$$

where $n = T-R$ and $\hat{\boldsymbol{\theta}}_{t-1}$ is consistently estimated from a rolling sample $\{y_{t-1},\ldots,y_{t-R}\}$ of size R. But we still do not know $f(\cdot)$. For this, we utilize the equivalence relationship between $\ln[f(y_t)/\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})]$ and the log-likelihood ratio of the inverse normal transform of the probability integral transform (PIT) of the actual realizations of the process with respect to the models' density forecast. This equivalence relationship enables us to consistently estimate $I(f:\psi,\boldsymbol{\theta}^*)$ and hence to conduct the out-of-sample comparison of possibly misspecified models in terms of their distance to the true model.

The (out-of-sample) PIT of the realization of the process with respect to the model's density forecast is defined as

$$u_t = \int_{-\infty}^{y_t}\psi(y;\hat{\boldsymbol{\theta}}_{t-1})\mathrm{d}y, \quad t = R+1,\ldots,T \tag{6}$$

It is well known that if $\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})$ coincides with the true density $f(y_t)$ for all t, then the sequence $\{u_t\}_{t=R+1}^{T}$ is IID and uniform on the interval [0, 1] ($U[0,1]$ henceforth). This provides a powerful approach to evaluating the quality of a density forecast model. Our task, however, is not to evaluate a single model, but to compare a battery of competing models. Our purpose of utilizing the PITs is to exploit the following equivalence between $\ln[f(y_t)/\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})]$ and the log likelihood ratio of the transformed PIT and hence to construct the distance measure. The inverse normal transform of the PIT is

$$x_t = \Phi^{-1}(u_t) \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. If the sequence $\{u_t\}_{t=R+1}^{T}$ is IID $U[0,1]$ then $\{x_t\}_{t=R+1}^{T}$ is IID standard normal $N(0,1)$ (IID $N(0,1)$ henceforth). More importantly, Berkowitz (2001, Proposition 2, p. 467) shows that

$$\ln\left[f(y_t)/\psi(y_t;\hat{\boldsymbol{\theta}}_{t-1})\right] = \ln\left[p(x_t)/\phi(x_t)\right] \tag{8}$$

where $p(\cdot)$ is the density of x_t and $\phi(\cdot)$ the standard normal density. Therefore, the distance of a density forecast model to the unknown true

model can be equivalently estimated by the departure of $\{x_t\}_{t=R+1}^{T}$ from IID $N(0, 1)$,

$$\tilde{I}(f : \psi) = \frac{1}{n}\sum_{t=R+1}^{T}[\ln p(x_t) - \ln \phi(x_t)] \tag{9}$$

We transform the departure of $\psi(\cdot;\boldsymbol{\theta})$ from $f(\cdot)$ to the departure of $p(\cdot)$ from IID $N(0, 1)$. To specify the departure from IID $N(0, 1)$, we want $p(\cdot)$ to be as flexible as possible to reflect the true distribution of $\{x_t\}_{t=R+1}^{T}$ and at the same time it can be IID $N(0, 1)$ if the density forecast model coincides with the true model. We follow Berkowitz (2001) by specifying $\{x_t\}_{t=R+1}^{T}$ as an AR(L) process

$$x_t = \boldsymbol{\rho}' X_{t-1} + \sigma\eta_t \tag{10}$$

where $X_{t-1} = (1, x_{t-1}, \ldots, x_{t-L})'$, $\boldsymbol{\rho}$ is an $(L+1)\times 1$ vector of parameters, and η_t IID distributed. We specify a flexible distribution for η_t, say, $p(\eta_t; \boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ is a vector of distribution parameters such that when $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ for some $\boldsymbol{\gamma}^*$ in the parameter space, $p(\eta_t; \boldsymbol{\gamma}^*)$ is IID $N(0, 1)$. A test for IID $N(0, 1)$ of $\{x_t\}_{t=R+1}^{T}$ per se can be constructed by testing elements of the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\rho}', \sigma, \boldsymbol{\gamma}')'$, say, $\boldsymbol{\rho} = \mathbf{0}$, $\sigma = 1$, and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. We assume the semi-nonparametric (SNP) density of order K of Gallant and Nychka (1987) for η_t

$$p(\eta_t; \boldsymbol{\gamma}) = \frac{\left(\sum_{k=0}^{K}\gamma_k \eta_t^k\right)\phi(\eta_t)}{\int_{-\infty}^{+\infty}\left(\sum_{k=0}^{K}\gamma_k u^k\right)^2 \phi(u)\mathrm{d}u}, \tag{11}$$

where $\gamma_0 = 1$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_K)'$. Setting $\gamma_k = 0$ $(k = 1, \ldots, K)$ gives $p(\eta_t) = \phi(\eta_t)$. Given (10) and (11), the density of x_t is

$$p(x_t; \boldsymbol{\beta}) = \frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma; \boldsymbol{\gamma}\right]}{\sigma},$$

which degenerates into IID $N(0,1)$ by setting $\boldsymbol{\beta} = \boldsymbol{\beta}^* = (\mathbf{0}', 1, \mathbf{0}')'$. Then $\tilde{I}(\varphi : \psi)$ as defined in (9) can be estimated by

$$\tilde{I}(f : \psi; \boldsymbol{\beta}) = \frac{1}{n}\sum_{t=R+1}^{T}\left[\ln\frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma; \boldsymbol{\gamma}\right]}{\sigma} - \ln\phi(x_t)\right].$$

The likelihood ratio test statistic of the adequacy of the density forecast model $\psi(\cdot;\boldsymbol{\theta})$ in Berkowitz (2001) is simply the above formula with $p(\cdot) = \phi(\cdot)$. As $\boldsymbol{\beta}$ is unknown, we estimate $\tilde{I}(\varphi : \psi)$ by

$$\tilde{I}(f : \psi; \hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{t=R+1}^{T}\left[\ln\frac{p\left[(x_t - \hat{\boldsymbol{\rho}}_n' X_{t-1})/\hat{\sigma}_n; \hat{\boldsymbol{\gamma}}_n\right]}{\hat{\sigma}_n} - \ln\phi(x_t)\right], \tag{12}$$

where $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\rho}}_n', \hat{\sigma}_n, \boldsymbol{\gamma}_n')'$ is the maximum likelihood estimator that maximizes $n^{-1}\sum_{t=R+1}^{T} \ln p(x_t; \boldsymbol{\beta})$.

To check the performance of a density forecast model in certain regions of the distribution, we can easily modify our distance measure tailored for the tail parts only. For the lower (left) tail, we define the censored random variable

$$x_t^\tau = \begin{cases} x_t & \text{if } x_t < \tau \\ \Phi^{-1}(\alpha) \equiv \tau & \text{if } x_t \geq \tau. \end{cases} \tag{13}$$

For example, $\tau = -1.645$ for $\alpha = 0.05$, the left 5% tail. As before, we consider an AR model (10) with η_t distributed as in (11). Then the censored random variable x_t^τ has the distribution function

$$p^\tau(x_t^\tau; \boldsymbol{\beta}) = \left[\frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma\right]}{\sigma}\right]^{\mathbf{1}(x_t < \tau)} \left[1 - P\left(\frac{\tau - \boldsymbol{\rho}' X_{t-1}}{\sigma}; \boldsymbol{\gamma}\right)\right]^{\mathbf{1}(x_t \geq \tau)}, \tag{14}$$

in which $P(\cdot; \boldsymbol{\gamma})$ is the CDF of the SNP density function and is calculated in the way as discussed in BLS. Given $p^\tau(x_t^\tau; \boldsymbol{\beta})$, the (left) tail minimum KLIC divergence measure can be estimated analogously

$$\tilde{I}^\tau(f : \psi; \hat{\boldsymbol{\beta}}_n^\tau) = \frac{1}{n}\sum_{t=R+1}^{T} \left[\ln p^\tau(x_t^\tau; \hat{\boldsymbol{\beta}}_n^\tau) - \ln \phi^\tau(x_t^\tau)\right], \tag{15}$$

where $\phi^\tau(x_t^\tau) = [1 - \Phi(\tau)]^{\mathbf{1}(x_t \geq \tau)}[\phi(x_t)]^{\mathbf{1}(x_t < \tau)}$ and $\hat{\boldsymbol{\beta}}_n^\tau$ maximizes $n^{-1}\sum_{t=R+1}^{T} \ln p^\tau(x_t^\tau; \boldsymbol{\beta})$.

For the upper (right) tail we define the censored random variable

$$x_t^\tau = \begin{cases} \Phi^{-1}(\alpha) \equiv \tau & \text{if } x_t \leq \tau \\ x_t & \text{if } x_t > \tau \end{cases} \tag{16}$$

For example, $\tau = 1.645$ for $\alpha = 0.95$, the right 5% tail. Then the censored random variable x_t^τ has the distribution function

$$p^\tau(x_t^\tau; \boldsymbol{\beta}) = \left[1 - P\left(\frac{\tau - \boldsymbol{\rho}' X_{t-1}}{\sigma}; \boldsymbol{\gamma}\right)\right]^{\mathbf{1}(x_t \leq \tau)} \left[\frac{p\left[(x_t - \boldsymbol{\rho}' X_{t-1})/\sigma\right]}{\sigma}\right]^{\mathbf{1}(x_t > \tau)}, \tag{17}$$

and the (right) tail minimum KLIC divergence measure can be estimated by (15) with $p^\tau(x_t^\tau; \boldsymbol{\beta})$ given by (17).

Therefore, given (6) and (7), we are able to estimate the minimum distance measure (4) by (12) (or its tail counterpart by (15)), which is proposed by BLS as a loss function to compare the out-of-sample predictive abilities of a set of $l+1$ competing models, each of which can be possibly misspecified. To

establish the notation with model indexed by k ($k = 0, 1, \ldots, l$), let the density forecast model k be denoted by $\psi_k(y_t;\ \boldsymbol{\theta}_k)$. Model comparison between a single model (model k) and the benchmark model (model 0) can be conveniently formulated as hypothesis testing of some suitable moment conditions. Consider constructing the loss differential

$$d_k = d_k(\psi_0, \psi_k) = \left[\ln f(y_t) - \ln\psi_0(y_t; \boldsymbol{\theta}_0^*)\right] - \left[\ln f(y_t) - \ln\psi_k(y_t; \boldsymbol{\theta}_k^*)\right] \quad (18)$$

where $1 \le k \le l$. Note that $\mathbb{E}(d_k) = I(f : \psi_0; \boldsymbol{\theta}_0^*) - I(f : \psi_k; \boldsymbol{\theta}_k^*)$ is the difference in the minimum KLIC of models 0 and k. When we compare multiple l models against a benchmark jointly, the null hypothesis of interest is that the best model is no better than the benchmark

$$\mathbb{H}_0 : \max_{1\le k\le l} \mathbb{E}(d_k) \le 0 \quad (19)$$

To implement this test, we follow White (2000) to bootstrap the following test statistic

$$\bar{V}_n \equiv \max_{1\le k\le l} \sqrt{n}\left[\bar{d}_{k,n} - \mathbb{E}(d_k)\right] \quad (20)$$

where $\bar{d}_{k,n} = \tilde{I}(f : \psi_0; \hat{\boldsymbol{\beta}}_{0,n}) - \tilde{I}(f : \psi_k; \hat{\boldsymbol{\beta}}_{k,n})$, and $\tilde{I}(f : \psi_0; \hat{\boldsymbol{\beta}}_{0,n})$ and $\tilde{I}(f : \psi_k; \hat{\boldsymbol{\beta}}_{k,n})$ are estimated by (12) for models 0 and k, with the normal-inversed PIT $\{x_t\}_{t=R+1}^{T}$ constructed using $\hat{\boldsymbol{\theta}}_{0,t-1}$ and $\hat{\boldsymbol{\theta}}_{k,t-1}$ estimated by a rolling-sample scheme, respectively. A merit of using the KLIC-based loss function for comparing forecasting models is that $\bar{d}_{k,n}$ incorporates both model specification error and parameter estimation error (note that x_t is constructed using $\hat{\boldsymbol{\theta}}$ rather than $\boldsymbol{\theta}^*$) as argued by BLS (2004).

To obtain the p-value for $\bar{V}_n$ White (2000) suggests using the stationary bootstrap of Politis and Romano (1994). This bootstrap p-value for testing $\mathbb{H}_0$ is called the "reality check p-value" for data snooping. As discussed in Hansen (2001), White's reality check p-value may be considered as an upper bound of the true p-value, since it sets $\mathbb{E}(d_k) = 0$. Hansen (2001) considers a modified reality check test, which removes the "bad" models from the comparison and thereby improves the size and the power of the test. The reality check to compare the performance of density forecast models in the tails can be implemented analogously.

4.2. Results of the BLS Test

We split the sample used in Section 2 into two parts (roughly into two halves): one for in-sample estimation of size $R = 1703$ and another for out-of-sample density forecast of size $n = 1700$. We use a rolling-sample scheme.

That is, the first density forecast is based on observations 1 through R (January 3, 1990–September 24, 1996), the second density forecast is based on observations 2 through $R+1$ (January 4, 1990–September 25, 1996), and so on.

The results are presented in Tables 2–4, with each table computing the statistics with different ways of selecting L and K. We present the reality check results with the whole density (100% with $\alpha = 1.00$), three left tails (10%, 5%, 1% with $\alpha = 0.10, 0.05, 0.01$), and three right tails (10%, 5%, 1% with $\alpha = 0.90, 0.95, 0.99$). With the AR($L$)–SNP($K$) model as specified in (10) and (11), we need to choose L and K. In Table 2, we fix $L = 3$ and $K = 5$. We minimize the Akaike information criteria (AIC) in Table 3, and the Schwarz information criteria (SIC) in Table 4, for the selection of L and K from the sets of $\{0, 1, 2, 3\}$ for L and $\{0, 1, \ldots, 8\}$ for K.

The out-of-sample average KLIC loss (denoted as "Loss" in tables) as well as the reality check p-value of White (2000) (denoted as "p_1") and the modified reality check p-value of Hansen (2001) (denoted as "p_2") are presented in these tables. In comparing the models using the reality check tests, we regard each model as a benchmark model and it is compared with the remaining eight models. We set the number of bootstraps for the reality check to be 1,000 and the mean block length to be 4 for the stationary bootstrap of Politis and Romano (1994). The estimated out-of-sample KLIC (12) and its censored versions as defined from (15) with different values of τ (each corresponding to different α) are reported in Tables 2–4. Note that in these tables a small value of the out-of-sample average loss and a large reality check p-value indicate that the corresponding model is a good density forecast model, as we fail to reject the null hypothesis that the other remaining eight models is no better than that model.

As our purpose is to test for the MDS property of the S&P 500 returns in terms of out-of-sample forecasts, we examine the performance of the MDS model as the benchmark (Model 0 with $k = 0$) in comparison with the remaining eight models indexed by $k = 1, \ldots, l$ ($l = 8$). The eight competing models are Constant, MA, ARMA, AR, PN, NN, SETAR, and STAR.

Table 2 shows the BLS test results computed using $L = 3$ and $K = 5$. First, comparing the entire return density forecasts with $\alpha = 100\%$, the KLIC loss value for the MDS model is $\tilde{I}(f : \psi_0; \hat{\boldsymbol{\beta}}_{0,n}) = 0.0132$, that is the smallest loss, much smaller than those of the other eight models. The large reality check p-values for the MDS model as the benchmark ($p_1 = 0.982$ and $p_2 = 0.852$) indicate that none of the other eight models are better than the MDS model, confirming the efficient market hypothesis that the S&P 500 returns may not be predictable using linear or nonlinear models.

Table 2. Reality Check Results Based on AR(3)-SNP(5).

Tail	Model	Left Tail					Right Tail				
		Loss	p_1	p_2	L	K	Loss	p_1	p_2	L	K
	MDS	0.0132	0.982	0.852	3	5					
	Constant	0.0233	0.370	0.370	3	5					
	MA	0.0237	0.357	0.357	3	5					
	ARMA	0.0238	0.357	0.357	3	5					
100%	AR	0.0171	0.547	0.486	3	5					
	PN	0.0346	0.071	0.071	3	5					
	NN	0.0239	0.355	0.355	3	5					
	SETAR	0.0243	0.354	0.354	3	5					
	STAR	0.0238	0.354	0.354	3	5					
	MDS	0.0419	1.000	0.730	3	5	0.0409	0.000	0.000	3	5
	Constant	0.0467	0.000	0.000	3	5	0.0359	0.415	0.394	3	5
	MA	0.0472	0.000	0.000	3	5	0.0352	0.794	0.686	3	5
	ARMA	0.0468	0.000	0.000	3	5	0.0359	0.442	0.397	3	5
10%	AR	0.0469	0.000	0.000	3	5	0.0363	0.309	0.298	3	5
	PN	0.0466	0.000	0.000	3	5	0.0348	0.627	0.615	3	5
	NN	0.0469	0.000	0.000	3	5	0.0365	0.269	0.261	3	5
	SETAR	0.0476	0.000	0.000	3	5	0.0360	0.396	0.372	3	5
	STAR	0.0470	0.000	0.000	3	5	0.0366	0.243	0.238	3	5
	MDS	0.0199	1.000	0.671	3	5	0.0229	0.000	0.000	3	5
	Constant	0.0218	0.002	0.002	3	5	0.0205	0.040	0.040	3	5
	MA	0.0217	0.003	0.003	3	5	0.0206	0.039	0.039	3	5
	ARMA	0.0217	0.003	0.003	3	5	0.0206	0.040	0.040	3	5
5%	AR	0.0211	0.023	0.000	3	5	0.0208	0.031	0.031	3	5
	PN	0.0226	0.000	0.000	3	5	0.0170	0.981	0.518	2	5
	NN	0.0211	0.026	0.000	3	5	0.0210	0.022	0.022	3	5
	SETAR	0.0211	0.040	0.004	3	5	0.0213	0.019	0.019	3	5
	STAR	0.0213	0.017	0.000	3	5	0.0208	0.029	0.029	3	5
	MDS	0.0068	1.000	0.992	3	5	0.0073	0.042	0.042	3	5
	Constant	0.0074	0.148	0.148	3	5	0.0068	0.172	0.172	3	5
	MA	0.0074	0.163	0.163	3	5	0.0068	0.167	0.167	3	5
	ARMA	0.0074	0.178	0.178	3	5	0.0068	0.160	0.160	3	5
1%	AR	0.0075	0.132	0.132	3	5	0.0068	0.151	0.151	3	5
	PN	0.0074	0.240	0.240	3	5	0.0058	0.932	0.914	3	5
	NN	0.0075	0.128	0.128	3	5	0.0068	0.150	0.150	3	5
	SETAR	0.0076	0.064	0.064	3	5	0.0068	0.135	0.135	3	5
	STAR	0.0074	0.181	0.181	3	5	0.0068	0.169	0.169	3	5

Note: "Loss" refers to is the out-of-sample averaged loss based on the KLIC measure; "p_1" and "p_2" are the reality check p-values of White's (2000) and Hansen's (2001) tests, respectively, where each model is regarded as a benchmark model and is compared with the remaining eight models. We use an AR(3)–SNP(5) model for the transformed PIT $\{x_t\}$. We retrieve the S&P 500 returns series from finance.yahoo.com. The sample observations are from January 3, 1990 to June 30, 2003 ($T = 3303$), the in-sample observations are from January 3, 1990 to September 24, 1996 ($R = 1703$), and the out-of-sample observations are from September 25, 1996 to June 30, 2003 ($n = 1700$).

Table 3. Reality Check Results Based on Minimum AIC.

Tail	Model	Left Tail					Right Tail				
		Loss	p_1	p_2	L	K	Loss	p_1	p_2	L	K
	MDS	0.0247	0.719	0.719	3	8					
	Constant	0.0231	0.938	0.933	3	3					
	MA	0.0266	0.534	0.534	3	7					
	ARMA	0.0236	0.874	0.862	3	3					
100%	AR	0.0266	0.533	0.533	3	7					
	PN	0.0374	0.238	0.238	1	8					
	NN	0.0238	0.853	0.845	3	3					
	SETAR	0.0266	0.539	0.539	3	7					
	STAR	0.0268	0.532	0.532	3	8					
	MDS	0.0583	1.000	0.767	3	8	0.0658	0.050	0.050	3	8
	Constant	0.0652	0.064	0.064	2	8	0.0593	0.717	0.675	3	8
	MA	0.0651	0.070	0.070	3	8	0.0582	0.960	0.947	3	8
	ARMA	0.0647	0.078	0.078	3	8	0.0588	0.839	0.813	3	8
10%	AR	0.0649	0.075	0.075	3	8	0.0594	0.701	0.655	3	8
	PN	0.0650	0.076	0.076	2	8	0.0628	0.239	0.239	3	8
	NN	0.0649	0.075	0.075	3	8	0.0597	0.619	0.571	3	8
	SETAR	0.0653	0.068	0.068	3	8	0.0584	0.845	0.808	3	8
	STAR	0.0653	0.063	0.063	3	8	0.0591	0.775	0.737	3	8
	MDS	0.0315	1.000	0.973	2	8	0.0390	0.333	0.092	3	8
	Constant	0.0334	0.628	0.473	2	7	0.0359	0.852	0.665	3	8
	MA	0.0339	0.468	0.320	2	8	0.0359	0.917	0.840	3	8
	ARMA	0.0340	0.425	0.266	3	8	0.0359	0.926	0.865	3	8
5%	AR	0.0336	0.495	0.340	3	8	0.0360	0.929	0.885	3	8
	PN	0.0356	0.256	0.129	3	8	0.0515	0.006	0.006	2	7
	NN	0.0337	0.487	0.327	3	8	0.0362	0.841	0.741	3	8
	SETAR	0.0334	0.529	0.386	3	8	0.0357	0.913	0.753	3	8
	STAR	0.0705	0.000	0.000	1	8	0.0359	0.908	0.826	3	8
	MDS	0.0098	0.973	0.939	3	7	0.0122	0.008	0.008	3	8
	Constant	0.0113	0.321	0.111	2	7	0.0113	0.042	0.042	3	8
	MA	0.0114	0.322	0.131	2	7	0.0116	0.018	0.018	3	8
	ARMA	0.0114	0.330	0.137	2	7	0.0118	0.019	0.019	3	8
1%	AR	0.0116	0.279	0.108	2	7	0.0118	0.020	0.020	3	8
	PN	0.0116	0.292	0.109	2	7	0.0092	0.989	0.564	3	7
	NN	0.0116	0.279	0.107	2	7	0.0118	0.020	0.020	3	8
	SETAR	0.0113	0.314	0.109	2	7	0.0112	0.060	0.060	3	7
	STAR	0.0229	0.000	0.000	2	8	0.0117	0.022	0.022	3	8

Note: "Loss" refers to is the out-of-sample averaged loss based on the KLIC measure; "p_1" and "p_2" are the reality check p-values of White's (2000) and Hansen's (2001) tests, respectively, where each model is regarded as a benchmark model and is compared with the remaining eight models; "L" and "K" are the AR and SNP orders, respectively, chosen by the minimum AIC criterion in the AR(L)–SNP(K) models, $L = 0, \ldots, 3$, $K = 0, \ldots, 8$ for the transformed PIT $\{x_t\}$. We retrieve the S&P 500 returns series from finance.yahoo.com. The sample observations are from January 3, 1990 to June 30, 2003 ($T = 3303$), the in-sample observations are from January 3, 1990 to September 24, 1996 ($R = 1703$), and the out-of-sample observations are from September 25, 1996 to June 30, 2003 ($n = 1700$).

Table 4. Reality Check Results Based on Minimum SIC.

Tail	Model	Left Tail					Right Tail				
		Loss	p_1	p_2	L	K	Loss	p_1	p_2	L	K
	MDS	0.0196	0.950	0.950	1	3					
	Constant	0.0211	0.628	0.628	1	3					
	MA	0.0216	0.562	0.562	1	3					
	ARMA	0.0215	0.570	0.570	1	3					
100%	AR	0.0215	0.580	0.580	1	3					
	PN	0.0342	0.157	0.157	2	4					
	NN	0.0214	0.599	0.599	1	3					
	SETAR	0.0221	0.538	0.538	1	3					
	STAR	0.0213	0.596	0.596	1	3					
	MDS	0.0583	1.000	0.767	3	8	0.0658	0.050	0.050	3	8
	Constant	0.0652	0.064	0.064	2	8	0.0593	0.717	0.675	3	8
	MA	0.0651	0.070	0.070	3	8	0.0582	0.960	0.947	3	8
	ARMA	0.0647	0.078	0.078	3	8	0.0588	0.839	0.813	3	8
10%	AR	0.0649	0.075	0.075	3	8	0.0594	0.701	0.655	3	8
	PN	0.0650	0.076	0.076	2	8	0.0628	0.239	0.239	3	8
	NN	0.0649	0.075	0.075	3	8	0.0597	0.619	0.571	3	8
	SETAR	0.0653	0.068	0.068	3	8	0.0584	0.845	0.808	3	8
	STAR	0.0653	0.063	0.063	3	8	0.0591	0.775	0.737	3	8
	MDS	0.0306	0.991	0.871	2	7	0.0390	0.333	0.092	3	8
	Constant	0.0334	0.384	0.213	2	7	0.0359	0.852	0.665	3	8
	MA	0.0328	0.416	0.241	3	7	0.0359	0.917	0.840	3	8
	ARMA	0.0328	0.411	0.238	3	7	0.0359	0.926	0.865	3	8
5%	AR	0.0324	0.495	0.318	3	7	0.0360	0.929	0.885	3	8
	PN	0.0343	0.250	0.098	3	7	0.0515	0.006	0.006	2	7
	NN	0.0325	0.489	0.311	3	7	0.0362	0.841	0.741	3	8
	SETAR	0.0323	0.522	0.357	3	7	0.0357	0.913	0.753	3	8
	STAR	0.0705	0.000	0.000	1	8	0.0359	0.908	0.826	3	8
	MDS	0.0000	1.000	0.544	3	1	0.0016	0.077	0.077	3	2
	Constant	0.0043	0.000	0.000	3	3	0.0014	0.090	0.090	3	2
	MA	0.0044	0.000	0.000	3	3	0.0000	1.000	0.999	3	1
	ARMA	0.0044	0.000	0.000	3	3	0.0000	1.000	0.999	3	1
1%	AR	0.0045	0.000	0.000	3	3	0.0000	1.000	0.997	3	1
	PN	0.0044	0.000	0.000	3	3	0.0014	0.109	0.109	3	2
	NN	0.0045	0.000	0.000	3	3	0.0000	0.852	0.718	3	1
	SETAR	0.0045	0.000	0.000	3	3	0.0015	0.095	0.095	3	2
	STAR	0.0229	0.000	0.000	2	8	0.0000	0.910	0.757	3	1

Note: "Loss" refers to is the out-of-sample averaged loss based on the KLIC measure; "p_1" and "p_2" are the reality checks p-values of White's (2000) and Hansen's (2001) tests, respectively, where each model is regarded as a benchmark model and is compared with the remaining eight models; "L" and "K" are the AR and SNP orders, respectively, chosen by the minimum SIC criterion in the AR(L)–SNP(K) models, $L = 0, \ldots, 3$, $K = 0, \ldots, 8$, for the transformed PIT $\{x_t\}$. We retrieve the S&P 500 returns series from finance.yahoo.com. The sample observations are from January 3, 1990 to June 30, 2003 ($T = 3303$), the in-sample observations are from January 3, 1990 to September 24, 1996 ($R = 1703$), and the out-of-sample observations are from September 25, 1996 to June 30, 2003 ($n = 1700$).

Next, comparing the left tails with $\alpha = 10\%, 5\%, 1\%$, we find the results are similar to those for the entire distribution. That is, the MDS model has the smallest KLIC loss values for these left tails, much smaller than those of the other eight models. The reality check p-values for the MDS model as the benchmark are all very close to one, indicating that none of the eight models are better than the MDS, confirming the efficient market hypothesis in the left tails of the S&P 500 returns.

Interestingly and somewhat surprisingly, when we look at the right tails with $\alpha = 90\%, 95\%, 99\%$ (i.e., the right 10%, 5%, and 1% tails, respectively), the results are exactly the opposite to those for the left tails. That is, the KLIC loss values for the MDS model for all of these three right tails are the largest, larger than those of the other eight competing models. The reality check p-values with the MDS model as the benchmark are zero or very close to zero, indicating that some of the other eight models are significantly better than the MDS model, hence rejecting the efficient market hypothesis in the right tails of the S&P 500 return density. This implies that the S&P 500 returns may be more predictable when the market goes up than when it goes down, during the sample period from January 3, 1990 to June 30, 2003. To our knowledge, this empirical finding is new to the literature, obtained as a benefit of using the BLS method that permits evaluation and comparison of the density forecasts on a particular area of the return density.

As the right tail results imply, the S&P 500 returns in the right tails are predictable via some of the eight linear and nonlinear competing models considered in this paper. To see the nature of the nonlinearity in mean, we compare the KLIC loss values of these models. It can be seen that the PN model has the smallest loss values for all the three right tails. The reality check p-values show that the PN model (as a benchmark) is not dominated by any of the remaining models for the three 10%, 5%, and 1% right tails. The other three nonlinear models (NN, SETAR, and STAR) are often worse than the linear models in terms of out-of-sample density forecasts. We note that, while PN is the best model for the right tails, it is the worst model for forecasting the entire density ($\alpha = 100\%$).

Summing up, the significant in-sample evidence from the generalized spectral statistics $M(1, 2)$ and $M(1, 4)$ reported in Table 1 suggests that the squared lagged return and the fourth order power of the lagged returns (i.e. the conditional kurtosis representing the influence of the tail observations) have a predictive power for the returns. The out-of-sample evidence adds that this predictability of the squared lagged return and the fourth order power of the lagged returns is *asymmetric*, i.e., significant only when the market goes up.

Table 3 shows the BLS test results computed with L and K chosen by the AIC. The results of Table 3 are similar to those of Table 2. Comparing the entire distribution with $\alpha = 100\%$, the large reality check p-values for the MDS model as the benchmark ($p_1 = 0.719$ and $p_2 = 0.719$) indicate that none of the other eight models are better than the MDS model (although the KLIC loss value for the MDS model is not the smallest). This confirms the market efficiency hypothesis that the S&P500 returns may not be predictable using linear or nonlinear models. The results for the left tails with $\alpha = 10\%$, 5%, 1% are also comparable to those in Table 2. That is, the KLIC loss values for the MDS model for these left tails are the smallest. The reality check p-values with the MDS model as the benchmark are all close to one ($p_1 = 1.000$, 1.000, 0.973 and $p_2 = 0.767$, 0.973, 0.939), indicating that none of the other eight models are better than the MDS model, again confirming the market efficiency hypothesis in the left tails of the S&P 500 returns. The right tail results are different from those for the left tails, as in Table 2. The MDS model for the right tails is generally worse than the other eight models. The reality check p-values of Hansen (2001) for the MDS model as the benchmark are very small ($p_2 = 0.050$, 0.092 and 0.008) for the three right tails, indicating that some of the eight models are significantly better than the MDS model. This shows that the S&P 500 returns may be more predictable when the market goes up than when it goes down – the nonlinear predictability is asymmetric.

Table 4 shows the BLS test results computed with L and K chosen by the SIC. The results are very similar to those of Tables 2 and 3. Comparing the entire distribution with $\alpha = 100\%$, the KLIC loss value for the benchmark MDS model is the smallest with the large reality check p-values ($p_1 = 0.950$ and $p_2 = 0.950$). The results for the left tails with $\alpha = 10\%$, 5%, 1% are consistent with those in Tables 2 and 3. That is, the KLIC loss values for the MDS model for these left tails are the smallest. The reality check p-values with the MDS as the benchmark are all close to one ($p_1 = 1.000$, 0.991, 1.000 and $p_2 = 0.767$, 0.871, 0.544), indicating that none of the other eight models are better than the MDS model, again confirming the market efficiency hypothesis in the left tails. The story for the right tails is different from that for the left tails, as in Tables 2 and 3. The MDS model for the right tails is generally worse than the other eight models. The reality check p-values of Hansen (2001) for the MDS model as the benchmark are very small ($p_2 = 0.050$, 0.092, and 0.077) for the three right tails, indicating that some of the linear and nonlinear models are significantly better than the MDS model. This shows that the S&P 500 returns may be more predictable when the market goes up than when it goes down.

5. CONCLUSIONS

In this paper, we examine nonlinearity in the conditional mean of the daily closing S&P 500 returns. We first examine the nature of the nonlinearity using the in-sample test of Hong (1999), where it is found that there are strong nonlinear predictable components in the S&P 500 returns. We then explore the out-of-sample nonlinear predictive ability of various linear and nonlinear models. The evidence for out-of-sample nonlinear predictability is quite weak in the literature, and it is generally accepted that stock returns follow a martingale. While most papers in the literature use MSFE, MAFE, or some economic measures (e.g., wealth or returns) to evaluate nonlinear models, we use the density forecast approach in this paper.

We find that for the entire distribution the S&P 500 daily closing returns are not predictable, and various nonlinear models we examine are no better than the MDS model. For the left tails, the returns are not predictable. The MDS model is the best in the density forecast comparison. For the right tails, however, the S&P 500 daily closing returns are found to be predictable by using some linear and nonlinear models. Hence, the out-of-sample nonlinear predictability of the S&P 1500 daily closing returns is asymmetric. These findings are robust to the choice of L and K in computation of the BLS statistics.

We note that the asymmetry in the nonlinear predictability which we have found is with regards to the two tails of the return distribution. Our results may not imply that the bull market is more predictable than the bear market because stock prices can fall (left tail) or rise (right tail) under both market conditions. Nonlinear models that incorporate the asymmetric tail behavior as well as the bull and bear market regimes would be an interesting model to examine, which we leave for future research.

NOTES

1. We would like to thank Yongmiao Hong for sharing his GAUSS code for the generalized spectral tests.
2. The GAUSS code is available upon request.

ACKNOWLEDGEMENT

We would like to thank the co-editor Tom Fomby and an anonymous referee for helpful comments. All remaining errors are our own.

REFERENCES

Anderson, T. G., Benzoni, L., & Lund, J. (2002). An empirical investigation of continuous-time equity return models. *Journal of Finance*, *57*, 1239–1284.

Bao, Y., Lee, T.-H., & Saltoglu, B. (2004). A test for density forecast comparison with applications to risk management. Working paper, University of California, Riverside.

Berkowitz, J. (2001). Testing density forecasts with applications to risk management. *Journal of Business and Economic Statistics*, *19*, 465–474.

Bradley, M. D., & Jansen, D. W. (2004). Forecasting with a nonlinear dynamic model of stock returns and industrial production. *International Journal of Forecasting*, *20*, 321–342.

Clements, M. P., & Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting*, *19*, 255–276.

Clements, M. P., & Smith, J. (2001). Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance*, *20*, 133–148.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*, 863–883.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*, 363–390.

Granger, C. W. J., & Pesaran, M. H. (2000a). A decision theoretic approach to forecasting evaluation. In: W. S. Chan, W. K. Li & Howell Tong (Eds), *Statistics and finance: An interface*. London: Imperial College Press.

Granger, C. W. J., & Pesaran, M. H. (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*, 537–560.

Hansen, P. R. (2001). An unbiased and powerful test for superior predictive ability. Working paper, Stanford University.

Hong, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association*, *84*, 1201–1220.

Hong, Y., & Lee, T.-H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, *85*(4), 1048–1062.

Kanas, A. (2003). Non-linear forecasts of stock returns. *Journal of Forecasting*, *22*, 299–315.

Kullback, L., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, *1*, 41–66.

Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, *3*, 175–208.

McMillan, D. G. (2001). Nonlinear predictability of stock market returns: Evidence from nonparametric and threshold Models. *International Review of Economics and Finance*, *10*, 353–368.

Meese, R. A., & Rogoff, K. (1983). The out of sample failure of empirical exchange rate models: Sampling error or misspecification. In: J. Frenkel (Ed.), *Exchange rates and international economics*. Chicago: University of Chicago Press.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*, 1303–1313.

Racine, J. (2001). On the nonlinear predictability of stock returns using financial and economic variables. *Journal of Business and Economic Statistics*, *19*(3), 380–382.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

Wright, J. H. (2000). Alternative variance-ratio tests using ranks and signs. *Journal of Business and Economic Statistics*, *18*, 1–9.

FLEXIBLE SEASONAL TIME SERIES MODELS

Zongwu Cai and Rong Chen

ABSTRACT

In this article, we propose a new class of flexible seasonal time series models to characterize the trend and seasonal variations. The proposed model consists of a common trend function over periods and additive individual trend (seasonal effect) functions that are specific to each season within periods. A local linear approach is developed to estimate the trend and seasonal effect functions. The consistency and asymptotic normality of the proposed estimators, together with a consistent estimator of the asymptotic variance, are obtained under the α-mixing conditions and without specifying the error distribution. The proposed methodologies are illustrated with a simulated example and two economic and financial time series, which exhibit nonlinear and nonstationary behavior.

1. INTRODUCTION

The analysis of seasonal time series has a long history that can be traced back to Gilbart (1854, 1856, 1865). Most of the time series, particularly economic and business time series, consist of trends and seasonal effects and many of them are nonlinear (e.g., Hylleberg, 1992; Franses, 1996, 1998).

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 63–87
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20022-1

There are basically three major types of trend and seasonality studied in the literature, and each requires different modeling approaches. When the trend and seasonality are determined to be stochastic, differences and seasonal differences of proper order are often taken to transform the time series to stationarity, which is then modeled by standard (seasonal) autoregressive and moving average (ARMA) models (e.g., Box, Jenkins, & Reinsel, 1994; Shumway & Stoffer, 2000; Pena, Tiao, & Tsay, 2001). This method is commonly used to model economic and financial data, such as the quarterly earnings data (e.g., Griffin, 1977; Shumway, 1988; Burman & Shumway, 1998; Franses, 1996, 1998). On the other hand, when the trend and seasonality are determined to be deterministic, linear (or nonlinear) additive or multiplicative trends and seasonal components are used, sometimes accompanied with an irregular noise component. More detailed discussion on these models can be found in the books by Shumway (1988), Brockwell and Davis (1991), and Franses (1996, 1998), and references therein. A third approach is the stable seasonal pattern model studied by Marshall and Oliver (1970, 1979), Chang and Fyffe (1971), and Chen and Fomby (1999). It assumes that given the season total, the probability portion of each period within the season remains constant across different periods. For more types of trend and seasonality studied in the literature, the reader is referred to the books by Hylleberg (1992), Franses (1996, 1998), and Ghysels and Osborn (2001).

In this article, we take the deterministic trend and seasonality approach. It is often observed, particularly in economic time series, that a seasonal pattern is clearly present but the seasonal effects change from period to period in some relatively smooth fashion. In addition, in many cases the trend cannot be fitted well by straight lines or even higher order polynomials because the characteristics of smooth variation change over time. These phenomena can be easily seen from the quarterly earnings series of Johnson & Johnson from 1960 to 1980. The data are taken from Table 7 of Shumway (1988, p. 358). Figure 1(a) shows the time series plot of the series. Clearly, there is an overall exponentially increasing trend with several different rates and a regular yearly seasonal component increasing in amplitude. A stochastic approach would first take a logarithm transformation (Fig. 1(b)), then a first order difference (Fig. 1(c)) and finally fit a conventional or seasonal autoregressive integrated moving-average model. But as pointed out by Burman and Shumway (1998), this approach does not seem to be satisfactory for this series due to its high nonlinearity and non-stationarity. It can be seen from Fig. 1(d), which displays the time series plot for each quarter over years, that the trends for all seasons are similar, increasing

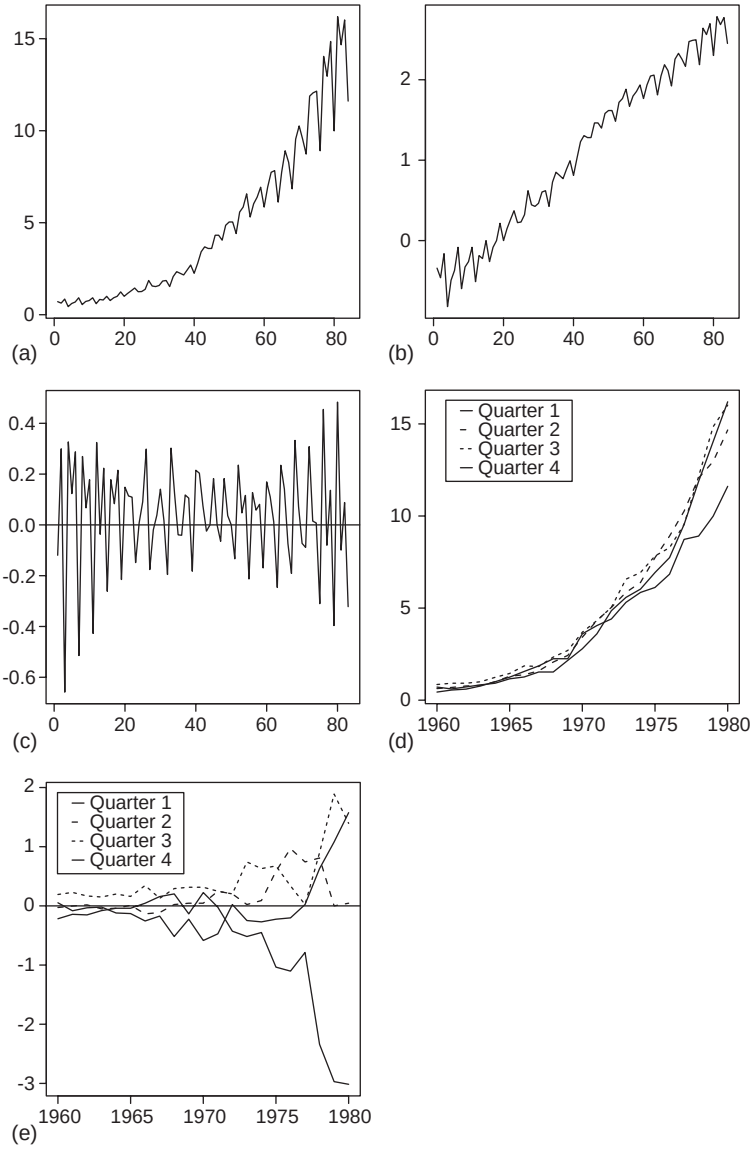


Fig. 1. Johnson & Johnson Quarterly Earnings Data from 1960–80. (a) Original. (b) Log-transformed Series. (c) Difference of the Log-transformed Series. (d) Quarter-by-Year Time Series Plots with the Legend: Solid Line for Quarter 1, Dotted Line for Quarter 2, Dashed Line for Quarter 3, and Dotted-Dashed Line for Quarter 4. (e) Quarter-by-Year Time Series Plots for the Observed Seasonal Data with the Same Legend as in (d).

smoothly but nonlinearly. In this article, we propose a flexible model that is capable of capturing such features in a time series.

Denote a seasonal time series as

$$y_{t1}, \ldots, y_{td}, \qquad t = 1, 2, \ldots,$$

where d is the number of seasons within a period. A general deterministic trend and seasonal component model assumes the form

$$y_{tj} = T_t + S_{tj} + e_{tj} \tag{1}$$

where T_t is the common trend same to different periods within a season, and S_{tj} is the seasonal effect, satisfying $\sum_{j=1}^{d} S_{tj} = 0$. A standard parametric model assumes a parametric function for the common trend T_t, such as linear or polynomial functions. The seasonal effects are usually assumed to be the same for different periods, that is, $S_{tj} = S_j$ for $j = 1, \ldots, d$ and all t, only if seasonality is postulated to be deterministic, an assumption which has been criticized in the seasonality literature.

Motivated by the Johnson & Johnson quarterly earnings data, Burman and Shumway (1998) considered the following semi-parametric seasonal time series model

$$y_{tj} = \alpha(t) + \beta(t)\gamma_j + e_{tj}, \qquad t = 1, \ldots, n, \qquad j = 1, \ldots, d \tag{2}$$

where $\alpha(t)$ is regarded as a common trend component and $\{\gamma_j\}$ are seasonal factors. Hence, the overall seasonal effect changes over periods in accordance with the modulating function $\beta(t)$. Implicitly, model (2) assumes that the seasonal effect curves have the same shape (up to a multiplicative constant) for all seasons. This assumption may not hold for some cases, as shown in the Johnson & Johnson series (Fig. 1(e)).

To estimate the functions and parameters in model (2), Burman and Shumway (1998) proposed two methods: nonlinear ANOVA and smoothing spline. They established the consistency for the estimators of trend, modulation function, and seasonal factors and the asymptotic normality for the estimators of the seasonal factors.

Inspired by the Johnson & Johnson quarterly earnings data of which the seasonal effect curves do not have the same shape over years (see Fig. 1(e)), we propose a more general flexible seasonal effect model having the following form:

$$y_{ij} = \alpha(t_i) + \beta_j(t_i) + e_{ij}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, d \tag{3}$$

where $t_i = i/n$, $\alpha(\cdot)$ is a smooth trend function in [0, 1], $\{\beta_j(\cdot)\}$ are smooth seasonal effect functions in [0, 1], either fixed or random, subject to a set of

constraints, and the error term e_{ij} is assumed to be stationary and strong mixing. Note that for convenience in study of the asymptotic properties, we use t_i for time index in (3), instead of the integer t. As in model (2), the following constraints are needed for fixed seasonal effects:

$$\sum_{j=1}^{d} \beta_j(t) = 0 \qquad \text{for all } t \tag{4}$$

reflecting the fact that the sum of all seasons should be zero for the seasonal factor. For random seasonal effects, $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_d(t))'$ is the vector of random seasonal effects with $E(\mathbf{1}'_d(\boldsymbol{\beta}(t)) = 0$ and a certain covariance structure, where, throughout, $\mathbf{1}_d$ denotes a $d \times 1$ unit vector. However, in this article, our focus is only on the fixed effect case and the stationary errors. Of course, it is warranted to further investigate the random effect case and it would be a very interesting topic to consider cases with nonstationary errors such as the integrated processes and stationary/nonstationary exogenous variables.

The rest of this article is organized as follows. In Section 2, we study in detail the flexible seasonal time series model (3). A local linear technique is used to estimate the trend and seasonal functions, and the asymptotic properties of the resulting estimators are studied. Section 3 is devoted to a Monte Carlo simulation study and the analysis of two economic and financial time series data. All the technical proofs are given in the appendix.

2. MODELING PROCEDURES

Combination of (3) and (4) in a matrix expression leads to

$$\mathbf{Y}_i = \mathbf{A}\boldsymbol{\theta}(t_i) + \mathbf{e}_i \tag{5}$$

where

$$\mathbf{Y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{id} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{1}_{d-1} & \mathbf{I}_{d-1} \\ 1 & -\mathbf{1}'_{d-1} \end{pmatrix}, \quad \boldsymbol{\theta}(t) = \begin{pmatrix} \alpha(t) \\ \beta_1(t) \\ \vdots \\ \beta_{d-1}(t) \end{pmatrix}, \quad \text{and} \quad \mathbf{e}_i = \begin{pmatrix} e_{i1} \\ \vdots \\ e_{id} \end{pmatrix}$$

$\mathbf{I}_d$ is the $d \times d$ identity matrix, and the error term $\mathbf{e}_i$ is assumed to be stationary with $E(\mathbf{e}_i) = 0$ and $\text{cov}(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{R}(i-j)$. Clearly, model (5) includes model (2) as a special case.

2.1. Local Linear Estimation

For estimating $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$ in (3), a local linear method is employed, although a general local polynomial method is also applicable. Local (polynomial) linear methods have been widely used in nonparametric regression during recent years due to their attractive mathematical efficiency, bias reduction, and adaptation of edge effects (e.g., Fan & Gijbels, 1996). Assuming throughout that $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$ have a continuous second derivative in [0, 1], then $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$ can be approximated by a linear function at any fixed time point $0 \le t \le 1$ as follows:

$$\alpha(t_i) \simeq a_0 + b_0(t_i - t) \quad \text{and} \quad \beta_j(t_i) \simeq a_j + b_j(t_i - t), \quad 1 \le j \le d-1,$$

where $\simeq$ denotes the first order Taylor approximation. Hence $\boldsymbol{\theta}(t_i) \simeq \mathbf{a} + \mathbf{b}(t_i - t)$, where $\mathbf{a} = \boldsymbol{\theta}(t)$ and $\mathbf{b} = \boldsymbol{\theta}^{(1)}(t) = d\boldsymbol{\theta}(t)/dt$ and (5) is approximated by

$$\mathbf{Y}_i \simeq \mathbf{Z}_i \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} + \mathbf{e}_i$$

where $\mathbf{Z}_i = (\mathbf{A},\ (t_i - t)\ \mathbf{A})$. Therefore, the locally weighted sum of least squares is

$$\sum_{i=1}^{n} \left\{ \mathbf{Y}_i - \mathbf{Z}_i \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \right\}' \left\{ \mathbf{Y}_i - \mathbf{Z}_i \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \right\} K_h(t_i - t) \tag{6}$$

where $K_h(u) = K(u/h)/h$, $K(\cdot)$ is the kernel function, and $h = h_n > 0$ is the bandwidth satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$, which controls the amount of smoothing used in the estimation.

By minimizing (6) with respect to $\mathbf{a}$ and $\mathbf{b}$, we obtain the local linear estimate of $\boldsymbol{\theta}(t)$, denoted by $\hat{\boldsymbol{\theta}}(t)$, which is $\hat{\mathbf{a}}$. It is easy to show that the minimizers of (6) are given by

$$\begin{pmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} S_{n,0}(t)\mathbf{A} & S_{n,1}(t)\mathbf{A} \\ S_{n,1}(t)\mathbf{A} & S_{n,2}(t)\mathbf{A} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{T}_{n,0}(t) \\ \mathbf{T}_{n,1}(t) \end{pmatrix}$$

where for $k \ge 0$,

$$S_{n,k}(t) = n^{-1} \sum_{i=1}^{n} (t_i - t)^k K_h(t_i - t) \quad \text{and} \quad \mathbf{T}_{n,k}(t) = n^{-1} \sum_{i=1}^{n} (t_i - t)^k K_h(t_i - t) \mathbf{Y}_i$$

Hence, the local linear estimator $\hat{\boldsymbol{\theta}}(t)$ is given by

$$\hat{\boldsymbol{\theta}}(t) = \mathbf{A}^{-1} \frac{S_{n,2}(t)\mathbf{T}_{n,0}(t) - S_{n,1}(t)\mathbf{T}_{n,1}(t)}{S_{n,0}(t)\, S_{n,2}(t) - S_{n,1}^2(t)} = \mathbf{A}^{-1} \sum_{i=1}^{n} S_i(t)\mathbf{Y}_i \tag{7}$$

where

$$S_i(t) = \frac{[S_{n,2}(t) - S_{n,1}(t)(t_i - t)]K_h(t_i - t)}{n\left\{S_{n,0}(t)S_{n,2}(t) - S_{n,1}^2(t)\right\}}$$

Remark 1. The local weighted least squares estimator from (6) does not take into the account of correlation between the periods $\mathbf{e}_t$ and $\mathbf{e}_s$. It can be improved by incorporating the correlation between periods. For example, given a known correlation structure, we can replace (6) with

$$\left\{\mathbf{Y} - \mathbf{Z}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right\}' \mathbf{K}_h^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{K}_h^{1/2}\left\{\mathbf{Y} - \mathbf{Z}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right\}$$

where $\mathbf{Y}_{nd\times 1}$ and $\mathbf{Z}_{nd\times 2d}$ are obtained by stacking $\mathbf{Y}_i$ and $\mathbf{Z}_i$, respectively, and Σ is the covariance matrix of stacked $\mathbf{e}_i$'s, and $\mathbf{K}_h^{1/2}$ is a diagonal matrix with $(id+j)$-th diagonal element being $K_h^{1/2}(t_i - t)(1 \le i \le n,\ 1 \le j \le d-1)$. When the correlation between the periods are unknown but can be modeled by certain parametric models such as the ARMA models, it is then possible to use an iterative procedure to estimate jointly the nonparametric function and the unknown parameters. In this article, we adopt the simple locally weighted least squares approach and use Eq. (6) to construct our estimator.

Remark 2. Note that many other nonparametric smoothing methods can be used here. The locally weighted least squares method is just one of the choices. There is a vast literature in theory and empirical study on the comparison of different methods (e.g., Fan & Gijbels, 1996).

Remark 3. The restriction to the locally weighted least squares method suggests that normality is at least being considered as a baseline. However, when non-normality is clearly present, a robust approach would be considered; see Cai and Ould-Said (2003) for details about this aspect in nonparametric regression estimation for time series.

Remark 4. The expression (7) has a nice interpretation which eases computation. Let $\mu_j(t) = \alpha(t) + \beta_j(t)$ be the seasonal mean curve for the jth season. Then $\hat{\mu}_j(t)$, the local linear estimator based on the series $y_{1j}, y_{2j}, \ldots, y_{nj}$

(and ignore all other observations), can be easily shown to be

$$\hat{\mu}_j(t) = \sum_{i=1}^{n} S_i(t) y_{ij}$$

which, in conjunction with (7), implies that

$$\hat{\boldsymbol{\theta}}(t) = \mathbf{A}^{-1} \hat{\boldsymbol{\mu}}(t)$$

where $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \ldots, \hat{\mu}_d(t))'$. A computation of the inverse of $\mathbf{A}$ leads to

$$\hat{\alpha}(t) = \frac{1}{d} \sum_{j=1}^{d} \hat{\mu}_j(t), \quad \text{and} \quad \hat{\beta}_j(t) = \hat{\mu}_j(t) - \hat{\alpha}(t), \quad 1 \le j \le d \tag{8}$$

Hence one needs to estimate only the low dimension individual mean curves $\{\hat{\mu}_j(t)\}$ to obtain $\hat{\alpha}(t)$ and $\{\hat{\beta}_j(t)\}$, which makes the implementations of the estimator much easier. However, we must point out that this feature is unique to the local linear estimator for model (5) and does not apply to other models. It is clear from (8) that $\{\mu_j(\cdot)\}$ can be estimated using different bandwidths, which can deal with the situation in which the seasonal mean functions have different degrees of smoothness.

Remark 5. Bandwidth selection is always one of the most important parts of any nonparametric procedure. There are several bandwidth selectors in the literature, including the generalized cross-validation of Wahba (1977), the plug-in method of Jones, Marron, and Sheather (1996), and the empirical bias method of Ruppert (1997), among others. They all can be used here. A comparison of different procedures can be found in Jones, Marron, and Sheather (1996). In this article, we use a procedure proposed in Fan, Yao, and Cai (2003), which combines the generalized cross-validation and the empirical bias method.

2.2. Asymptotic Theory

The estimation method described in Section 2.1 can accommodate both fixed and random designs. Here, our focus is on fixed design. The reason is that it might be suitable for pure time series data, such as financial and economic data. Since data are observed in time order as in Burman and Shumway (1998, p. 137), we assume that $t_i = t_{ni} = i/n$ for simplicity although the theoretical results developed later still hold for non-equal spaced design points, and that $\mathbf{e}_i = \mathbf{e}_{ni}$. In such a case, we assume that, for each n, $\{\mathbf{e}_{n1}, \ldots, \mathbf{e}_{nn}\}$ have the same joint distribution as $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$, where $\boldsymbol{\xi}_t$,

$t = \ldots,-1, 0, 1, \ldots$, is a strictly stationary time series defined on a probability space $(\Omega, \mathscr{A}, \mathscr{P})$ and taking values on $\mathscr{R}^d$. This type of assumption is commonly used in fixed-design regression for time series contexts. Detailed discussions on this respect can be found in Roussas (1989), Roussas, Tran, and Ioannides (1992), and Tran, Roussas, Yakowitz, and Van (1996) for nonparametric regression estimation for dependent data.

Traditionally, the error component in a deterministic trend and seasonal component like model (1) is assumed to follow certain linear time series models such as an ARMA process. Here we consider a more general structure – the α-mixing process, which includes many linear and nonlinear time series models as special cases (see Remark 6). Our theoretical results here are derived under the α-mixing assumption. For reference convenience, the mixing coefficient is defined as

$$\alpha(i) = \sup\{|P(A \cap B) - P(A)P(B)| : \quad A \in \mathscr{F}_{-\infty}^{0}, \quad B \in \mathscr{F}_{i}^{\infty}\}$$

where $\mathscr{F}_{a}^{b}$ is the σ-algebra generated by $\{\xi_i\}_{i=a}^{b}$. If $\alpha(i) \to 0$ as $i \to \infty$, the process is called strongly mixing or α-mixing.

Remark 6. Among various mixing conditions used in the literature, α-mixing is reasonably weak and is known to be fulfilled for many linear and nonlinear time series models under some regularity conditions. Gorodetskii (1977) and Withers (1981) derived the conditions under which a linear process is α-mixing. In fact, under very mild assumptions, linear autoregressive and more generally bilinear time series models are α-mixing with mixing coefficients decaying exponentially. Auestad and Tjøstheim (1990) provided illuminating discussions on the role of α-mixing (including geometric ergodicity) for model identification in nonlinear time series analysis. Chen and Tsay (1993) showed that the functional autoregressive process is geometrically ergodic under certain conditions. Further, Masry and Tjøstheim (1995, 1997) and Lu (1998) demonstrated that under some mild conditions, both autoregressive conditional heteroscedastic processes and nonlinear additive autoregressive models with exogenous variables, particularly popular in finance and econometrics, are stationary and α-mixing. Roussas (1989) considered linear processes without satisfying the mixing condition. Potentially our results can be extended to such cases.

Assumptions:

A1. Assume that the kernel $K(u)$ is symmetric and satisfies the Lipschitz condition and $u\,K(u)$ is bounded, and that $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$ have a continuous second derivative in [0,1].

A2. For each n, $\{\mathbf{e}_{n1}, \ldots, \mathbf{e}_{nn}\}$ have the same joint distribution as $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$, where $\boldsymbol{\xi}_t$, $t = \ldots, -1, 0, 1, \ldots$, is a strictly stationary time series with the covariance matrix $\mathbf{R}(k-l) = \mathrm{cov}(\boldsymbol{\xi}_k, \boldsymbol{\xi}_l)$ defined on a probability space $(\Omega, \mathscr{A}, \mathscr{P})$ and taking values on $\Re d$.[1] Assume that the time series $\{\boldsymbol{\xi}_i\}$ is α-mixing with the finite $2(1+\delta)$th moment for some $\delta > 0$ (i.e. $\mathrm{E}|\boldsymbol{\xi}_i|^{2(1+\delta)} < \infty$) and the mixing coefficient $\alpha(n)$ satisfying $\alpha(n) = O(n^{-(2+\delta)(1+\delta)/\delta})$.

A3. $nh^{1+4/\delta} \to \infty$.

Remark 7. Let $r_{jm}(k)$ denote the (j, m)-th element of $\mathbf{R}(k)$. By the Davydov's inequality (e.g., Corollary A.2 in Hall and Heyde, 1980), Assumption A2 implies $\sum_{k=-\infty}^{\infty} |r_{jm}(k)| < \infty$.

Note that all the asymptotic results here assume that the number of periods $n \to \infty$ but the number of seasons within a period d is fixed. Define, for $k \geq 0$, $\mu_k = \int u^k K(u)\, du$ and $\upsilon_k = \int u^k K^2(u)\, d\, u$. Let $\boldsymbol{\Sigma}_0 = \sum_{k=-\infty}^{\infty} \mathbf{R}(k)$, which exists by Assumption A2 (see Remark 7), and $\boldsymbol{\theta}^{(2)}(t) = d^2\boldsymbol{\theta}(t)/dt^2$. Now we state the asymptotic properties of the local linear estimator $\hat{\boldsymbol{\theta}}(t)$. A sketch of the proofs is relegated to the appendix.

Theorem 1. Under Assumptions A1 and A2, we have

$$\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) - \frac{h^2}{2}\mu_2\boldsymbol{\theta}^{(2)}(t) + o(h^2) = O_p((n\,h)^{-1/2})$$

Theorem 2. Under Assumptions A1–A3, we have

$$\sqrt{n\,h}\left\{\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) - \frac{h^2}{2}\mu_2\boldsymbol{\theta}^{(2)} + o(h^2)\right\} \to N(\mathbf{0},\ \boldsymbol{\Sigma}_\theta)$$

where $\boldsymbol{\Sigma}_\theta = \upsilon_0\mathbf{A}^{-1}\boldsymbol{\Sigma}_0(\mathbf{A}^{-1})'$. Clearly, the asymptotic variance of $\hat{\alpha}(t)$ is $\upsilon_0 d^{-2}\mathbf{1}_d'\boldsymbol{\Sigma}_0\mathbf{1}_d$.

Remark 8. As a consequence of Theorem 1, $\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) = O_p(h^2 + (n\,h)^{-1/2})$ so that $\hat{\boldsymbol{\theta}}(t)$ is a consistent estimator of $\boldsymbol{\theta}(t)$. From Theorem 2, the asymptotic mean square error (AMSE) of $\hat{\boldsymbol{\theta}}(t)$ is given by

$$\mathrm{AMSE} = \frac{h^2}{4}\mu_2^2||\boldsymbol{\theta}^2(t)||_2^2 + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\theta)}{n\,h}$$

which gives the optimal bandwidth, $h_{opt} = n^{-1/5}\left\{\mathrm{tr}(\boldsymbol{\Sigma}_\theta\mu_2^{-2})||\boldsymbol{\theta}^{(2)}(t)||_2^{-2}\right\}^{-1/5}$, by minimizing the AMSE, where $||A||_2 = (\sum_i\sum_j a_{ij}^2)^{1/2}$ and $\mathrm{tr}(A) = \sum_i a_{ii}$ if $\mathrm{A} = (a_{ij})$ is a square matrix. Hence, the optimal convergent rate of the

AMSE for $\hat{\boldsymbol{\theta}}(t)$ is of the order of $n^{-4}/^{5}$, as one would have expected. Also, the asymptotic variance of the estimator does not depend on the time point t. More importantly, it shows that the asymptotic variance of the estimator depends on not only the covariance structure of the seasonal effects ($\mathbf{R}(0) = \text{var}(\mathbf{e}_i)$) but also the autocorrelations over periods ($\sum_{k=1}^{\infty}\mathbf{R}(k)$). Finally, it is interesting to note that in general, the elements (except the first one) in the first row of Σ_θ are not zero, unless $\Sigma_0 = \sigma^2\mathbf{I}$. This implies that $\hat{\alpha}(t)$ and $\hat{\beta}_j(t)$ $(1 \le j \le d-1)$ may be asymptotically correlated.

Remark 9. In practice, it is desirable to have a quick and easy implementation to estimate the asymptotic variance of $\hat{\boldsymbol{\theta}}(t)$ to construct a pointwise confidence interval. The explicit expression of the asymptotic variance in Theorem 2 provides two direct estimators. From Lemma A1 in the Appendix, for any $0 < t < 1$, we have

$$\boldsymbol{\Sigma}_\theta = \lim_{n\to\infty}\frac{h}{n}\mathbf{A}^{-1}\text{var}\left(\sum_{i=1}^{n}\mathbf{e}_{ni}K_h(t_i - t)\right)(\mathbf{A}^{-1})'$$

Hence a direct (naive) estimator of Σ_θ is given by $\widehat{\boldsymbol{\Sigma}}_\theta = \mathbf{A}^{-1}\widehat{\mathbf{Q}}_{n0}\widehat{\mathbf{Q}}'_{n0}(\mathbf{A}^{-1})'$, where $\widehat{\mathbf{Q}}_{n0} = (h\,n^{-1})^{1/2}\sum_{i=1}^{n}\{\mathbf{Y}_i - \mathbf{A}\hat{\boldsymbol{\theta}}(t_i)\}\,K_h(t_i - t)$. However, in the finite sample, $\widehat{\boldsymbol{\Sigma}}_\theta$ might depend on t. To overcome this shortcoming, an alternative way to construct the estimation of $\boldsymbol{\Sigma}_0$ is to use some popular approach in econometric literature such as heteroskedasticity consistent (HC) or heteroskedasticity and autocorrelation consistent (HAC) method; see White (1980), Newey and West (1987, 1994).

Remark 10. Let $\widehat{\mathbf{Y}}(\tau)$ be the forecast of $\mathbf{Y}(\tau)$ based on the mean function. Then, $\widehat{\mathbf{Y}}(\tau) = \mathbf{A}\hat{\boldsymbol{\theta}}(\tau)$ and

$$\mathbf{Y}(\tau) - \hat{\mathbf{Y}}(\tau) = \mathbf{A}(\boldsymbol{\theta}(\tau) - \hat{\boldsymbol{\theta}}(\tau)) + \mathbf{e}_\tau$$

so that the forecast variance, by ignoring the correlation between $\hat{\boldsymbol{\theta}}(\tau)$ and $\mathbf{e}_\tau$ and the bias, is

$$\begin{aligned}&\text{var}\,(\mathbf{Y}(\tau) - \hat{\mathbf{Y}}(\tau)) \approx \mathbf{A}\,\text{var}\,(\boldsymbol{\theta}(\tau) - \hat{\boldsymbol{\theta}}(\tau))\mathbf{A}'\\ &+ \mathbf{R}(0) \approx \upsilon_0(n\,h)^{-1}\boldsymbol{\Sigma}_0 + \mathbf{R}(0)\end{aligned}$$

by Theorem 2. Therefore, the standard error estimates for the predictions can be taken to be the square root of the diagonal elements of $\upsilon_0(n\,h)^{-1}\hat{\boldsymbol{\Sigma}}_0 + \widehat{\mathbf{R}}(0)$, where $\hat{\boldsymbol{\Sigma}}_0$ is given in the above remark and $\widehat{\mathbf{R}}(0) = n^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{A}\,\hat{\boldsymbol{\theta}}(t_i))(\mathbf{Y}_i - \mathbf{A}\,\widehat{\boldsymbol{\theta}}(t_i))'$. Note that the above formula is just an approximation. A further theoretical and empirical study on the predictive utility based on model (3) is warranted.

3. EMPIRICAL STUDIES

In this section, we use a simulated example and two real examples to illustrate our proposed models and the estimation procedures. Throughout this section, we use the Epanechnikov kernel, $K(u) = 0.75(1 - u^2)I(|u| \leq 1)$ and the bandwidth selector mentioned in Remark 5.

Example 1. A Simulated Example. : We begin the illustration with a simulated example of the flexible seasonal time series model. For this simulated example, the performance of the estimators is evaluated by the mean absolute deviation error (MADE)

$$\varepsilon = n_0^{-1} \sum_{k=1}^{n_0} |\widehat{\alpha}(u_k) - \alpha(u_k)|$$

for $\alpha(\cdot)$ and

$$\varepsilon_j = n_0^{-1} \sum_{k=1}^{n_0} |\widehat{\beta}_j(u_k) - \beta_j(u_k)|$$

for $\beta_j(\cdot)$, where $\{u_k,\ k = 1, \ldots, n_0\}$ are the grid points from (0, 1).

In this simulated example, we consider the following nonlinear seasonal time series model

$$y_{ij} = \alpha(t_i) + \beta_j(t_i) + e_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, 4$$

where $t_i = i/n$

$$\begin{aligned}
\alpha(x) &= \exp(-0.7 + 3.5x) \\
\beta_1(x) &= -3.1x^2 + 17.1x^4 - 28.1x^5 + 15.7x^6 \\
\beta_2(x) &= -0.5x^2 + 15.7x^6 - 15.2x^7 \\
\beta_4(x) &= -0.2 + 4.8x^2 - 7.7x^3 \\
\beta_3(x) &= -\beta_1(x) - \beta_2(x) - \beta_4(x)
\end{aligned}$$

for $0 < x \leq 1$. Here, the errors $e_{ij} = e_{4(i-1)+j}$, where $\{e_t\}$ are generated from the following AR(1) model:

$$e_t = 0.9e_{t-1} + \varepsilon_t$$

and ε_t is generated from $N(0, 0.1^2)$. The simulation is repeated 500 times for each of the sample sizes $n = 50$, 100, and 300. Figure 2 shows the results of a typical example for sample size $n = 100$. The typical example is selected in such a way that its total MADE value $(= \varepsilon + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4)$ is equal to

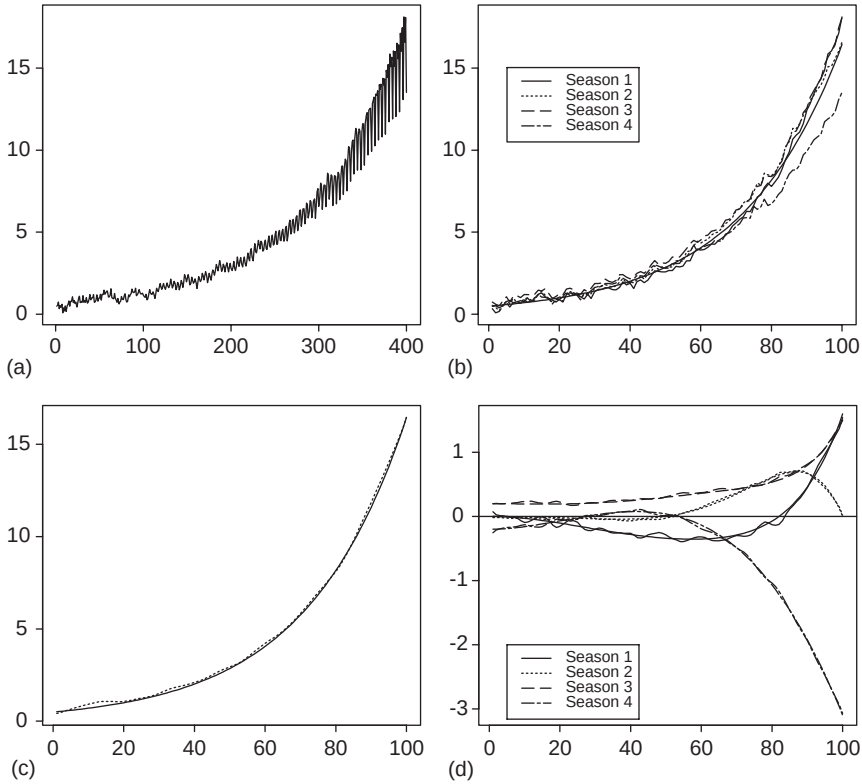


Fig. 2. Simulation Results for Example 1. (a) The Time Series Plot of the Time Series $\{y_t\}$. (b) The Time Series Plots for $\{y_{ij}\}$ for Each Season (Thin Lines) with True Trend Function $\alpha(\cdot)$ (Thick Solid Line) with the Legend: Solid Line for Season 1, Dotted Line for Season 2, Dashed Line for Season 3, and Dotted-Dashed Line for Season 4. (c) The Local Linear Estimator (Dotted Line) of the Trend Function $\alpha(\cdot)$ (Solid Line). (d) The Local Linear Estimator (Thin Lines) of the Seasonal Effect Functions $\{\beta_j(\cdot)\}$ (Thick Solid Lines) with the Legend: Solid Line for $\beta_1(\cdot)$, Dotted Line for $\beta_2(\cdot)$, Dashed Line for $\beta_3(\cdot)$, and Dotted-Dashed Line for $\beta_4(\cdot)$.

the median in the 500 replications. Figure 2(a) presents the time series $\{y_{(i-1)4+j}\}$ and Figure 2(b) gives the time series plots of each season $\{y_{ij}\}$ (thin lines) with the true trend function $\alpha(\cdot)$ (thick solid line). Figures 2(c) and 2(d) display, respectively, the estimated $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$ (thin lines), together with their true values (thick lines). They show that the estimated values are very close to the true values. The median and standard deviation (in parentheses) of the 500 MADE values are summarized in

Table 1. The Median and Standard Deviation of the 500 MADE Values.

n	ε	ε_1	ε_2	ε_3	ε_4
50	0.1202(0.0352)	0.0284(0.0077)	0.0271(0.0083)	0.0244(0.0071)	0.0263(0.0073)
100	0.0997(0.0286)	0.0204(0.0051)	0.0214(0.0059)	0.0179(0.0044)	0.0192(0.0057)
300	0.0692(0.0161)	0.0135(0.0031)	0.0151(0.0037)	0.0119(0.0027)	0.0122(0.0034)

Table 1, which shows that all the MADE values decrease as n increases. This reflects the asymptotic results. Overall, the proposed modeling procedure performs fairly well.

Example 2. Johnson & Johnson Earnings Series: We continue our study on the Johnson & Johnson quarterly earnings data, described in Section 1. There are two components of the quarterly earnings data: (1) a four-period seasonal component and (2) an adjacent quarter component which describes the seasonally adjusted series. As discussed in Section 1, we fit the series using

$$y_{ij} = \alpha(t_i) + \beta_j(t_i) + e_{ij}, \quad i = 1, \ldots, 21, \quad j = 1, \ldots, 4 \tag{9}$$

subject to the constraint $\sum_{j=1}^{4} \beta_j(t) = 0$ for all t. Together with the observed data for each quarter (thin lines) over years, the estimated trend function $\widehat{\alpha}(\cdot)$ (thick solid line) is plotted in Fig. 3(a), and the estimated seasonal effect functions $\{\hat{\beta}_j(\cdot)\}$ (thick lines) are displayed in Fig. 3(b) with the observed seasonal data

$$\bar{y}_{ij} = y_{ij} - \frac{1}{4}\sum_{j=1}^{4} y_{ij}$$

for each quarter (thin lines) over years. With the automatically selected bandwidth, the fitted and observed values are very close to each other. For comparison purpose, we calculate the overall mean squared error

$$\frac{1}{84}\sum_{i=1}^{21}\sum_{j=1}^{4}\left\{y_{ij} - \hat{\alpha}(t_i) - \hat{\beta}_j(t_i)\right\}^2$$

which is 0.0185, with the effective number of parameters (e.g., Hastie & Tibshirani, 1990; Cai & Tiwari, 2000) being 44.64. It seems to be a significant improvement of the semi-parametric model (2) of Burman and Shumway (1998), which has a MSE of 0.0834 with 44 effective number of parameters.

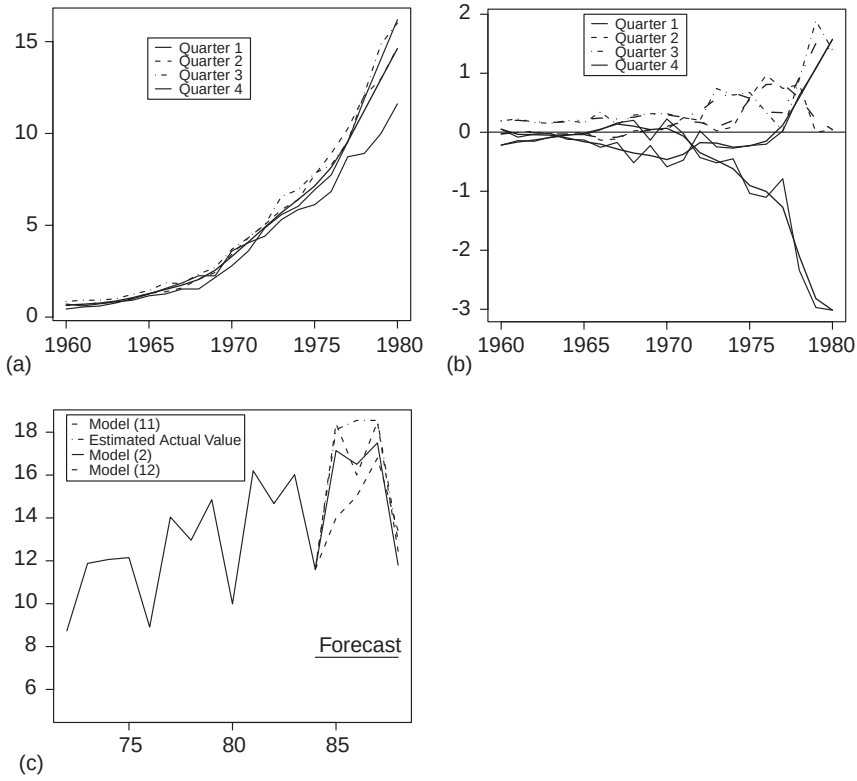


Fig. 3. Johnson & Johnson Quarterly Earnings Data from 1960–80. (a) Quarter-by-Year Time Series Plots (Thin Lines) with the Estimated Trend Functions (Thick Line) with the Same Legend: Solid Line for Quarter 1, Dotted Line for Quarter 2, Dashed Line for Quarter 3, and Dotted-Dashed Line for Quarter 4. (b) Quarter-by-Year Time Series Plots for the Observed Seasonal Data (Thin Lines) with the Estimated Seasonal Effect Functions (Thick Lines) with the Same Legend as in (a). (c) Four-Quarter Forecasts for 1981 Based on: Model (9), Dotted Line, Model (2), Dotted-Dashed Line, and Model (10), Short-Dashed Line, together with "Estimated Actual" Value – Dashed Line.

Finally, to evaluate the prediction performance, we use the whole dataset (all 84 quarters from 1960 to 1980) to fit the model and forecast the next four quarters (in 1981) by using one-step ahead prediction $\hat{y}_{22,j} = \hat{\alpha}(t_{22}) + \hat{\beta}_j(t_{22})$ for $1 \le j \le 4$ in model (9). The standard error estimates for the predictions (see Remark 10) are 0.110, 0.106, 0.186, and 0.177, respectively.

Also, we consider the predictions using the following seasonal regression model, popular in the seasonality literature (e.g., Hylleberg, 1992; Franses, 1996, 1998; Ghysels & Osborn, 2001), for the first order difference of the logarithm transformed data

$$z_t = \log(y_t) - \log(y_{t-1}) \sum_{j=1}^{4} \theta_j D_{j,t} + \varepsilon_t, \quad 2 \le t \le 84 \tag{10}$$

where $\{D_{j,t}\}$ are seasonal dummy variables. Figure 3(c) reports the predictions for the mean value function based on models (9) (dotted line) and (10) (short-dashed line), connected with the observed values (solid line) for the last thirteen quarters (from quarters 72 to 84). In contrast to the semi-parametric model proposed by Burman and Shumway (1998) which forecasts a slight decreasing bend in the trend function for 1981 (the dashed-dotted line), our model predicts an upward trend. Indeed, $\hat{\alpha}(t_{22}) = 16.330$, which is much larger than the last two values of the estimated trend function $\hat{\alpha}(t_{21}) = 14.625$ and $\hat{\alpha}(t_{20}) = 12.948$. For comparison purpose, in Fig. 3(c), together with the forecasting results based on model (2) from Burman and Shumway (1998) in dashed-dotted line and model (10) in short-dashed line, we show, in dashed line, the "estimated actual" earnings for the year 1981. They were "estimated" since we were not able to obtain the exact definition of the earnings of the original series and had to resort to estimating the actual earning (using the average) in the scale of the original data from various sources. Fig. 3(c) shows that the forecasting results based on model (9) is closer to the "estimated actual" earnings than those based on model (2) from Burman and Shumway (1998) and model (10), although neither forecasting result for the second quarter of 1981 is close to the "estimated actual" earning, and that model (10) does not predict well.

Example 3. US Retail Sales Series: In this example we apply the proposed methods to the monthly US retail sales series (not seasonally adjusted) from January of 1967 to December of 2000 (in billions of US dollars). The data can be downloaded from the web site at http://marketvector.com. The US retail sales index is one of the most important indicators of the US economy. There are vast studies of this series in the literature (e.g., Franses, 1996, 1998; Ghysels & Osborn, 2001). Figure 4(a) represents the series plotted as 12 curves, each corresponding to a specific month of each calendar year and the yearly averages plotted in the thick solid line. They show that the trends for all seasons (months) are similar, basically increasing but nonlinearly. Figure 4(b) displays the monthly growth over years and shows pattern of high seasonal effects. It also can be observed

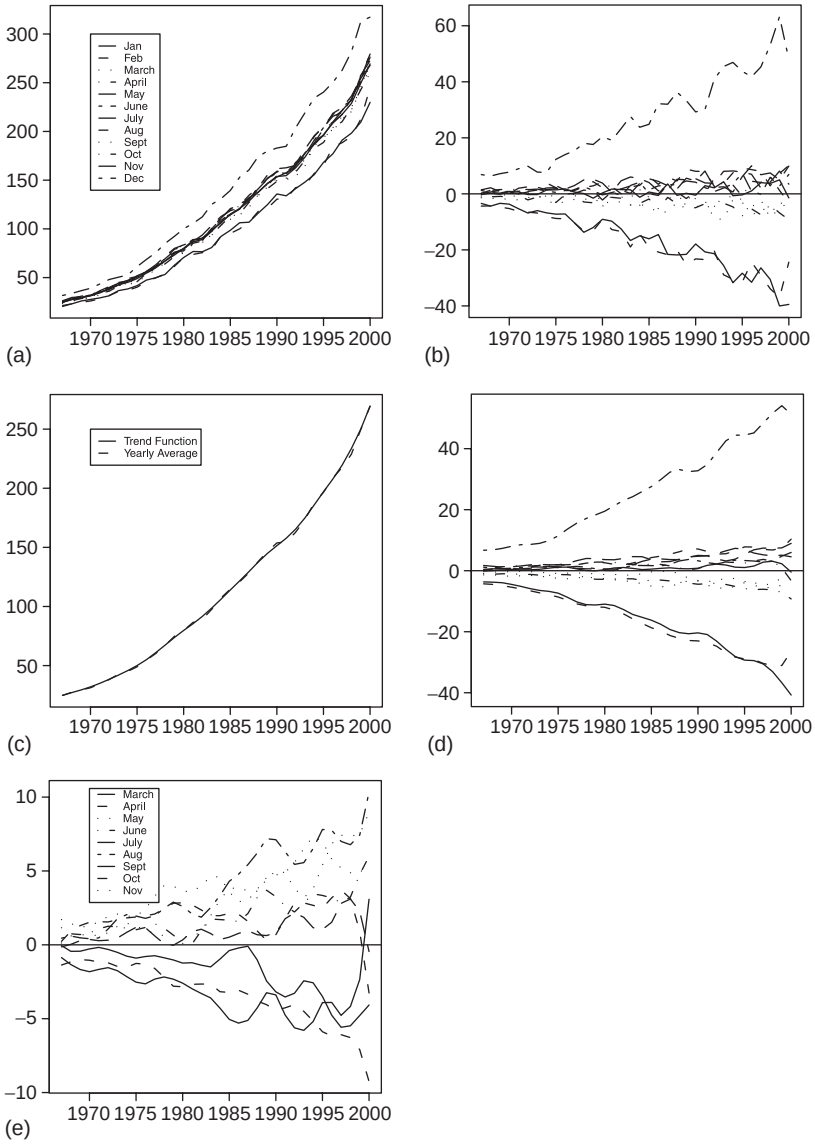


Fig. 4. US Retail Sales Data from 1967–2000. (a) Quarter-by-Year Time Series Plots with Yearly Averages (Thick Solid Line). (b) Quarter-by-Year Time Series Plots for the Observed Seasonal Data with the Same Legend as in (a). (c) Estimated Trend Function (Solid Line) with the Yearly Averages (Dashed Line). (d) Estimated Seasonal Functions with the Same Legend as in (a). (e) Estimated Seasonal Functions for Months from March to November.

that the sales were decreasing over the years for the first four months in the year and September, and increasing for the rest seven months. In particular, the decreasing/increasing rates for January, February, and December were much faster than those for the rest nine months. Retail sales are known to be large in the fourth quarter because of Christmas spending and small in the first quarter. This is confirmed from Fig. 4(b) where the monthly curve for December was significantly higher than the rest months and those for January and February were much lower than the rest.

We fit the series using the following nonparametric seasonal model

$$y_{ij} = \alpha(t_i) + \beta_j(t_i) + e_{ij}, \quad i = 1, \ldots, 34, \quad \text{and} \quad j = 1, \ldots, 12$$

with the constraints $\sum_{j=1}^{12} \beta_j(t_i) = 0$ for all t_i. The overall mean squared error

$$\frac{1}{408} \sum_{i=1}^{34} \sum_{j=1}^{12} \{y_{ij} - \hat{\alpha}(t_i) - \hat{\beta}_j(t_i)\}^2$$

is 4.64 and the correlation coefficient between the estimated data and the actual data is 0.9996. Figure 4(c) depicts the estimated trend function with the yearly averages (dotted line) and it shows that the trend is increasing nonlinearly (possibly exponentially). The estimated seasonal effect functions are plotted in Fig. 4(d) for all months. From these plots we observe a stable pattern – the seasonal effect functions for the first four months and September stay below the average and the rest of them are above the average. Also, the monthly curves drift apart and some of them cross each other. Particularly, the monthly curves for January, February, and December dominate those for the rest nine months. To get a better picture of the months from March to November, the estimated seasonal effect functions for these nine months are depicted in Fig. 4(e). It appears that some patterns change starting from 1999.

NOTE

1. This stationary assumption might not be suitable for many economic applications, and the models might allow the errors to be non-stationary such as an integrated process. This topic is still open and beyond the scope of this article, although, a further investigation is warranted.

ACKNOWLEDGMENTS

The authors are grateful to the co-editor T. Fomby and the referee for their very detailed and constructive commends and suggestions, and Professors D. Findley, T. Fomby, R.H. Shumway, and R.S. Tsay as well as the audiences at the NBER/NSF Time Series Conference at Dallas 2004 for their helpful and insightful comments and suggestions. Cai's research was supported in part by the National Science Foundation grants DMS-0072400 and DMS-0404954 and funds provided by the University of North Carolina at Charlotte. Chen's research was supported in part by the National Science Foundation grants DMS-0073601, 0244541 and NIH grant R01 Gm068958.

REFERENCES

Auestad, B., & Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika*, *77*, 669–687.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. G. (1994). *Time series analysis, forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Brockwell, P. J., & Davis, R. A. (1991). *Time series theory and methods*. New York: Springer.

Burman, P., & Shumway, R. H. (1998). Semiparametric modeling of seasonal time series. *Journal of Time Series Analysis*, *19*, 127–145.

Cai, Z., Fan, J., & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*, 941–956.

Cai, Z., & Ould-Said, E. (2003). Local M-estimator for nonparametric time series. *Statistics and Probability Letters*, *65*, 433–449.

Cai, Z., & Tiwari, R. C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, *11*, 341–350.

Chang, S. H., & Fyffe, D. E. (1971). Estimation of forecast errors for seasonal style-Goods sales. *Management Science*, *Series B 18*, 89–96.

Chen, R., & Fomby, T. (1999). Forecasting with stable seasonal pattern models with an application of Hawaiian tourist data. *Journal of Business & Economic Statistics*, *17*, 497–504.

Chen, R., & Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, *88*, 298–308.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.

Fan, J., Yao, Q., & Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of Royal Statistical Society*, *Series B 65*, 57–80.

Franses, P. H. (1996). *Periodicity and stochastic trends in economic time series*. New York: Cambridge University Press.

Franses, P. H. (1998). *Time series models for business and economic forecasting*. New York: Cambridge University Press.

Ghysels, E., & Osborn, D. R. (2001). *The econometric analysis of seasonal time series*. New York: Cambridge University Press.

Gilbart, J. W. (1854). The law of the currency as exemplified in the circulation of country bank notes in England since the passing of the Act of 1844. *Statistical Journal*, *17*.

Gilbart, J. W. (1856). The laws of currency in Scotland. *Statistical Journal*, *19*.

Gilbart, J. W. (1865). *Logic for the million, a familiar exposition of the art of reasoning*. London: Bell and Daldy.

Gorodetskii, V. V. (1977). On the strong mixing property for linear sequences. *Theory of Probability and Its Applications*, *22*, 411–413.

Griffin, P. A. (1977). The time series behavior of quarterly earnings: Preliminary evidence. *Journal of Accounting Research*, *15*, 71–83.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its applications*. New York: Academic Press.

Hastie, T. J., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman and Hall.

Hylleberg, S. (1992). The historical perspective. In: S. Hylleberg (Ed.), *Modelling Seasonality*. Oxford: Oxford University Press.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*, 401–407.

Lu, Z. (1998). On the ergodicity of non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*, *8*, 1205–1217.

Masry, E., & Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory*, *11*, 258–289.

Masry, E., & Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, *13*, 214–252.

Marshall, K. T., & Oliver, R. M. (1970). A constant work model for student enrollments and attendance. *Operations Research Journal*, *18*, 193–206.

Marshall, K. T., & Oliver, R. M. (1979). Estimating errors in student enrollment forecasting. *Research in Higher Education*, *11*, 195–205.

Newey, W. K., & West, K. D. (1987). A simple, positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, *61*, 631–653.

Pena, D., Tiao, G. C., & Tsay, R. S. (2001). *A course in time series analysis*. New York: Wiley.

Roussas, G. G. (1989). Consistent regression estimation with fixed design points under dependence conditions. *Statistics and Probability Letters*, *8*, 41–50.

Roussas, G. G., Tran, L. T., & Ioannides, D. A. (1992). Fixed design regression for time series: Asymptotic normality. *Journal of Multivariate Analysis*, *40*, 262–291.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, *92*, 1049–1062.

Shumway, R. H. (1988). *Applied statistical time series analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis & its applications*. New York: Springer-Verlag.

Tran, L., Roussas, G., Yakowitz, S., & Van, B. T. (1996). Fixed-design regression for linear time series. *The Annals of Statistics*, *24*, 975–991.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In: P. R. Krisnaiah (Ed.), *Applications of Statistics* (pp. 507–523). Amsterdam, North Holland.

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica*, *48*, 817838.

Withers, C. S. (1981). Conditions for linear processes to be strong mixing. *Zeitschrift fur Wahrscheinlichkeitstheorie verwandte Gebiete*, *57*, 477–480.

APPENDIX: PROOFS

Note that the same notations in Section 2 are used here. Throughout this appendix, we denote by C a generic constant, which may take different values at different appearances.

Lemma A1. Under the assumptions of Theorem 1, we have

$$\text{var}(\mathbf{Q}_{n0}) = \upsilon_0 \boldsymbol{\Sigma}_0 + o(1) \quad \text{and} \quad \mathbf{Q}_{n1} = o_p(1)$$

where for $k = 0, 1$,

$$\mathbf{Q}_{nk} = h^{1/2} n^{-1/2} \sum_{i=1}^{n} (t_i - t)^k \mathbf{e}_{ni} K_h(t_i - t)$$

Proof. By the stationarity of $\{\boldsymbol{\xi}_j\}$,

$$\begin{aligned}
\text{var}(\mathbf{Q}_{n0}) &= n^{-1} h \sum_{1 \le k,l \le n} \mathbf{R}(k-l) K_h(t_k - t) K_h(t_l - t) \\
&= n^{-1} h\, \mathbf{R}(0) \sum_{k=1}^{n} K_h^2(t_k - t) + 2n^{-1} h \sum_{1 \le l < k \le n} \mathbf{R}(k-l) K_h(t_k - t) K_h(t_l - t) \\
&\equiv \mathbf{I}_1 + \mathbf{I}_2
\end{aligned}$$

Clearly, by the Riemann sum approximation of an integral,

$$\mathbf{I}_1 \approx \mathbf{R}(0) h \int_0^1 K_h^2(u - t) du \approx \upsilon_0 \mathbf{R}(0)$$

Since $nh \to \infty$, there exists $c_n \to \infty$ such that $c_n/(nh) \to 0$. Let $S_1 = \{(k, l): 1 \le k - l \le c_n;\ 1 \le l < k \le n\}$ and $S_2 = \{(k, l) : 1 \le l < k \le n\} - S_1$. Then, $\mathbf{I}_2$ is split into two terms as $\sum_{S_1}(\cdots)$, denoted by $\mathbf{I}_{21}$, and $\sum_{S_2}(\cdots)$, denoted by $\mathbf{I}_{22}$. Since $K(\cdot)$ is bounded, then, $K_h(\cdot) \le C/h$ and $n^{-1} \sum_{k=1}^{n} K_h(t_k - t) \le C$. In conjunction with the Davydov's inequality (e.g., Corollary A.2 in

Hall and Heyde, 1980), we have, for the (j, m)-th element of $\mathbf{I}_{22}$,

$$\begin{aligned}
\left|\mathbf{I}_{22(jm)}\right| &\leq Cn^{-1}h\sum_{S_2}|r_{jm}(k-l)|K_h(t_k-t)(K_h(t_l-t)\\
&\leq Cn^{-1}h\sum_{S_2}\alpha^{\delta/(2+\delta)}(k-l)K_h(t_k-t)K_h(t_l-t)\\
&\leq Cn^{-1}\sum_{k=1}^{n}K_h(t_k-t)\sum_{k_1>c_n}\alpha^{\delta/(2+\delta)}(k_1)\\
&\leq C\sum_{k_1>c_n}\alpha^{\delta/(2+\delta)}(k_1)\leq Cc_n^{-\delta}\to 0
\end{aligned}$$

by Assumption A2 and the fact that $c_n\to\infty$. For any $(k, l)\in S_1$, by Assumption A1

$$|K_h(t_k-t)-K_h(t_l-t)|\leq Ch^{-1}(t_k-t_l)/h\leq Cc_n/(nh^2)$$

$$\begin{aligned}
\left|\mathbf{I}_{21(jm)}\right| &\equiv \left|2n^{-1}h\sum_{l=1}^{n-1}\sum_{1\leq k-l\leq c_n}r_{jm}(k-l)\{K_h(t_k-t)-K_h(t_l-t)\}K_h(t_l-t)\right|\\
&\leq Cc_nn^{-2}h^{-1}\sum_{l=1}^{n-1}\sum_{1\leq k-l\leq c_n}|r_{jm}(k-l)|K_h(t_l-t)\\
&\leq Cc_nn^{-2}h^{-1}\sum_{l=1}^{n-1}K_h(t_l-t)\sum_{k\geq 1}|r_{jm}(k)|\leq Cc_n/(n\,h)\to 0
\end{aligned}$$

by Remark 7 and the fact that $c_n/(nh)\to 0$. Therefore,

$$\begin{aligned}
\mathbf{I}_{21(jm)} &= 2n^{-1}h\sum_{l=1}^{n-1}\sum_{1\leq k-l\leq c_n}r_{jm}(k-l)K_h(t_k-t)K_h(t_l-t)\\
&= 2n^{-1}h\sum_{l=1}^{n-1}K_h^2(t_l-t)\sum_{1\leq k-l\leq c_n}r_{jm}(k-l)+\mathbf{I}_{212(jm)}\to 2\upsilon_0\sum_{k=1}^{\infty}r_{jm}(k)
\end{aligned}$$

Thus,

$$\text{var}\,(\mathbf{Q}_{n0})\to\upsilon_0\left(\mathbf{R}(0)+2\sum_{k=1}^{\infty}\mathbf{R}(k)\right)=\upsilon_0\boldsymbol{\Sigma}_0$$

On the other hand, by Assumption A1, we have

$$\begin{aligned}\text{var}\,(\mathbf{Q}_{n1}) &= n^{-1}h \sum_{1\le k,l\le n} \mathbf{R}(k-l)(t_k-t)(t_l-t)K_h(t_k-t)K_h(t_l-t) \\ &\le Cn^{-1}h \sum_{1\le k,l\le n} |\mathbf{R}(k-l)| \le C\,h \sum_{k=-\infty}^{\infty} |\mathbf{R}(k)| \to 0\end{aligned}$$

This proves the lemma.

Proof of Theorem 1. By the Riemann sum approximation of an integral,

$$S_{n,k}(t) = h^k \mu_k + o(1) \tag{A.1}$$

By the Taylor expansion, for any $k \ge 0$ and t_i in a neighborhood of t

$$\begin{aligned}&n^{-1} \sum_{i=1}^{n} (t_i - t)^k \boldsymbol{\theta}(t_i) K_h(t_i - t) \\ &= S_{n,k}(t)\boldsymbol{\theta}(t) + S_{n,k+1}(t)\boldsymbol{\theta}^{(1)}(t) + \frac{1}{2} S_{n,k+2}(t)\boldsymbol{\theta}^{(2)}(t) + o(h^2)\end{aligned}$$

so that by (7),

$$\begin{aligned}\hat{\boldsymbol{\theta}}(t) = \boldsymbol{\theta}(t) &+ \frac{1}{2} \frac{S_{n,2}^2(t) - S_{n,1}(t)S_{n,3}(t)}{S_{n,0}(t)S_{n,2}(t) - S_{n,1}^2(t)} \boldsymbol{\theta}^{(2)}(t) \\ &+ o(h^2) + \mathbf{A}^{-1} \sum_{i=1}^{n} S_i(t)\mathbf{e}_{ni}\end{aligned}$$

Then, by (A.1),

$$\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) - \frac{h^2}{2}\mu_2 \boldsymbol{\theta}^{(2)}(t) + o(h^2) = \mathbf{A}^{-1} \sum_{i=1}^{n} S_i(t)\mathbf{e}_{ni}$$

which implies that

$$\begin{aligned}&\sqrt{n\,h}\left\{\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) - \frac{h^2}{2}\mu_2 \partial^2 \boldsymbol{\theta}(t)/\partial t^2 + o(h^2)\right\} \\ &= \mathbf{A}^{-1} \frac{S_{n,2}(t)\,\mathbf{Q}_{n0} - s_{n,1}(t)\mathbf{Q}_{n1}}{S_{n,0}(t)S_{n,2}(t) - S_{n,1}^2(t)}\end{aligned} \tag{A.2}$$

where both $\mathbf{Q}_{n0}$ and $\mathbf{Q}_{n1}$ are defined in lemma A1. An application of lemma A1 proves the theorem.

Proof of Theorem 2. It follows from (A.2) that the term $\frac{1}{2}h^2\mu_2\partial^2\boldsymbol{\theta}(t)/\partial t^2$ on the right hand side of (A.2) serves as the asymptotic bias. Also, by (A.1) and lemma A1, we have

$$\sqrt{n\,h}\left\{\hat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}(t)-\frac{h^2}{2}\mu_2\partial^2\boldsymbol{\theta}(t)/\partial t^2+o(h^2)\right\}$$
$$=\mathbf{A}^{-1}\{1+o(1)\}\{\mathbf{Q}_{n0}+o_p(1)\}$$

which implies that to establish the asymptotic normality of $\hat{\boldsymbol{\theta}}(t)$, one only needs to consider the asymptotic normality for $\mathbf{Q}_{n0}$. To this end, the Cramér-Wold device is used. For any unit vector $\mathbf{d}\in\Re^d$, let $Z_{n,i}=n^{-1/2}h^{1/2}\,\mathbf{d}'\,\mathbf{e}_{ni}\,K_h(t_i-t)$. Then, $\mathbf{d}'\mathbf{Q}_{n0}=\sum_{i=1}^n Z_{n,i}$ and by lemma A1,

$$\text{var}\,(\mathbf{d}'\mathbf{Q}_{n0})=v_0\mathbf{d}'\boldsymbol{\Sigma}_0\mathbf{d}\{1+o(1)\}\equiv\theta_d^2\{1+o(1)\} \tag{A.3}$$

Now, the Doob's small-block and large-block technique is used. Namely, partition $\{1,\ldots,n\}$ into $2q_n+1$ subsets with large-block of size

$$r_n=\left\lfloor(nh)^{1/2}\right\rfloor$$

and small-block of size

$$s_n=\left\lfloor(nh)^{1/2}/\log\,n\right\rfloor,$$

where

$$q_n=\left\lfloor\frac{n}{r_n+s_n}\right\rfloor.$$

Then, $q_n\alpha(s_n)\le C\,n^{-(\tau-1)/2}\,h^{-(\tau+1)/2}\log^{\tau}n$, where $\tau=(2+\delta)(1+\delta)/\delta$, and $q_n\,\alpha(s_n)\to 0$ by Assumption A3. Let $r_j^*=j(r_n+s_n)$ and define the random variables, for $0\le j\le q_n-1$,

$$\eta_j=\sum_{i=r_j^*+1}^{r_j^*+r_n}Z_{n,i},\quad \xi_j=\sum_{i=r_j^*+r_n+1}^{r_{j+1}^*}Z_{n,i},\quad\text{and}\quad Q_{n,3}=\sum_{i=r_{qn}^*+1}^{n}Z_{n,i}$$

Then, $\mathbf{d}'\mathbf{Q}_{n0}=Q_{n,1}+Q_{\mathrm{n},2}+Q_{n,3}$, where $Q_{n-1}=\sum_{j=0}^{qn-1}\eta_j$ and $Q_{n,2}=\sum_{j=0}^{q_{n-1}}\xi_j$. Next we prove the following four facts: (i) as $n\to\infty$,

$$E\big(Q_{n,2}\big)^2\to 0,\quad E\big(Q_{n,3}\big)^2\to 0 \tag{A.4}$$

(ii) as $n \to \infty$ and $\theta_d^2(t)$ defined as in (A.3), we have

$$\sum_{j=0}^{q_n-1} E(\eta_j^2) \to \theta_d^2 \tag{A.5}$$

(iii) for any t and $n \to \infty$,

$$\left| E[\exp(itQ_{n,1})] - \prod_{j=0}^{q_n-1} E[\exp(it\eta_j)] \right| \to 0 \tag{A.6}$$

and (iv) for every $\varepsilon > 0$,

$$\sum_{j=0}^{q_n-1} E\left[\eta_j^2 I\left\{|\eta_j| \geq \varepsilon\theta_d\right\}\right] \to 0 \tag{A.7}$$

(A.4) implies that $Q_{n,2}$ and $Q_{n,3}$ are asymptotically negligible in probability. (A.6) shows that the summands $\{\eta_j\}$ in $Q_{n,1}$ are asymptotically independent, and (A.5) and (A.7) are the standard Lindeberg-Feller conditions for asymptotic normality of $Q_{n,1}$ for the independent setup. The rest proof is to establish (A.4)–(A.7) and it can be done by following the almost same lines as those used in the proof of Theorem 2 in Cai, Fan, and Yao (2000) with some modifications. This completes the proof of Theorem 2.

This page intentionally left blank

ESTIMATION OF LONG-MEMORY TIME SERIES MODELS: A SURVEY OF DIFFERENT LIKELIHOOD-BASED METHODS

Ngai Hang Chan and Wilfredo Palma

ABSTRACT

Since the seminal works by Granger and Joyeux (1980) and Hosking (1981), estimations of long-memory time series models have been receiving considerable attention and a number of parameter estimation procedures have been proposed. This paper gives an overview of this plethora of methodologies with special focus on likelihood-based techniques. Broadly speaking, likelihood-based techniques can be classified into the following categories: the exact maximum likelihood (ML) estimation (Sowell, 1992; Dahlhaus, 1989), ML estimates based on autoregressive approximations (Granger & Joyeux, 1980; Li & McLeod, 1986), Whittle estimates (Fox & Taqqu, 1986; Giraitis & Surgailis, 1990), Whittle estimates with autoregressive truncation (Beran, 1994a), approximate estimates based on the Durbin–Levinson algorithm (Haslett & Raftery, 1989), state-space-based maximum likelihood estimates for ARFIMA models (Chan & Palma, 1998), and estimation of stochastic volatility models (Ghysels, Harvey, & Renault, 1996; Breidt, Crato, & de Lima, 1998; Chan & Petris, 2000) among others. Given the diversified

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 89–121
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20023-3

applications of these techniques in different areas, this review aims at providing a succinct survey of these methodologies as well as an overview of important related problems such as the ML estimation with missing data (Palma & Chan, 1997), influence of subsets of observations on estimates and the estimation of seasonal long-memory models (Palma & Chan, 2005). Performances and asymptotic properties of these techniques are compared and examined. Inter-connections and finite sample performances among these procedures are studied. Finally, applications to financial time series of these methodologies are discussed.

1. INTRODUCTION

Long-range dependence has become a key aspect of time series modeling in a wide variety of disciplines including econometrics, hydrology and physics, among many others. Stationary long-memory processes are defined by autocorrelations decaying slowly to zero or spectral density displaying a pole at zero frequency. A well-known class of long-memory models is the autoregressive fractionally integrated moving average (ARFIMA) processes introduced by Granger and Joyeux (1980) and Hosking (1981). An ARFIMA process $\{y_t\}$ is defined by

$$\Phi(B)(1-B)^d y_t = \Theta(B)\varepsilon_t \tag{1}$$

where $\Phi(B) = 1 + \phi_1 B + \cdots + \phi_p B^p$ and $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ are the autoregressive and moving average (ARMA) operators, respectively; $(1-B)^d$ is the fractional differencing operator defined by the binomial expansion $(1-B)^d = \sum_{j=0}^{\infty} \binom{d}{j} B^j$ and $\{\varepsilon_t\}$ a white noise sequence with variance σ_ε^2. Under the assumption that the roots of the polynomials $\Phi(B)$ and Θ(B) are outside the unit circle and $|d| < 1/2$, the ARFIMA(p,d,q) process is second-order stationary and invertible. The spectral density of this process is

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} |1 - e^{i\omega}|^{-2d} |1 - \Phi(e^{i\omega})|^{-2} |1 - \Theta(e^{i\omega})|^2$$

and its ACF may be written as

$$\gamma(k) = \int_{-\pi}^{\pi} f(\omega) e^{i\omega k} d\omega \tag{2}$$

Estimation of long-memory models has been considered by a large number of authors. Broadly speaking, most of the estimation methodologies

proposed in the literature can be classified into the time-domain and the spectral-domain procedures. In the first group, we have the exact maximum likelihood estimators (MLE) and the quasi maximum likelihood estimators (QMLE), see for example the works by Granger and Joyeux (1980), Sowell (1992) and Beran (1994a), among others. In the second group, we have for instance, the Whittle and the semi-parametric estimators (see Fox & Taqqu, 1986; Giraitis & Surgailis, 1990; Robinson, 1995) among others.

In this article, we attempt to give an overview of this plethora of estimation methodologies, examining their advantages and disadvantages, computational aspects such as their arithmetic complexity, finite sample behavior and asymptotic properties.

The remaining of this paper is structured as follows. Exact ML methods are reviewed in Section 2, including the Cholesky decomposition, the Levinson–Durbin algorithm and state-space methodologies. Section 3 discusses approximate ML methods based on truncations of the infinite autoregressive (AR) expansion of the long-memory process, including the Haslett and Raftery estimator and the Beran method. Truncations of the infinite moving average (MA) expansion of the process are discussed in Section 4 along with the corresponding Kalman filter recursions. The spectrum-based Whittle method and semi-parametric procedures are studied in Section 5. Extensions to the non-Gaussian case are also addressed in this section. The problem of parameter estimation of time series with missing values is addressed in Section 6 along with an analysis of the effects of data gaps on the estimates. Section 7 discusses methodologies for estimating time series displaying both persistence and cyclical behavior. Long-memory models for financial time series are addressed in Section 8, while final remarks are presented in Section 9.

2. EXACT MAXIMUM LIKELIHOOD METHOD

Under the assumption that the process $\{y_t\}$ is Gaussian and has zero mean, the log-likelihood function may be written as

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2}\log\ \det\ \Gamma_\theta - \frac{1}{2}\, Y'\Gamma_\theta^{-1}\, Y \tag{3}$$

where $Y = (y_1, \ldots, y_n)'$, $\Gamma_\theta = \mathrm{var}(Y)$ and θ is the model parameter vector. Hence, the ML estimate is given by $\hat{\theta} = \mathrm{argmax}_\theta \mathscr{L}(\boldsymbol{\theta})$.

Expression (3) involves the calculation of the determinant and the inverse of Γ_θ. As described next, the well-known Cholesky decomposition algorithm

can be used to carry out these calculations. Further details can be found in Sowell (1992). Here we give an overview of this and other related methodologies for computing (3) such as the Levinson–Durbin algorithm and the Kalman filter recursions.

2.1. Cholesky Decomposition

Since the variance–covariance matrix Γ_θ is positive definite, it can be decomposed as

$$\Gamma_\theta = L_1 L_1'$$

where L_1 is a lower triangular matrix. This Cholesky decomposition provides the determinant $\det \Gamma_\theta = (\det L_1)^2 = \prod_{i=1}^n l_{ii}^2$, where l_{ii} denotes the ith-diagonal element of L_1. Furthermore, the inverse of Γ_θ can be obtained as $\Gamma_\theta^{-1} = (L_1^{-1})' L_1^{-1}$, where the inverse of L_1 can be computed by means of a very simple procedure, see for example Press, Teukolsky, Vetterling, and Flannery (1992, 89ff).

Observe that while the inversion of a nonsingular square matrix $n \times n$ has arithmetic complexity of order $\mathcal{O}(n^3)$, the Cholesky algorithm is of order $\mathcal{O}(n^3/6)$, cf. Press et al. (1992, p. 34).

2.2. Levinson–Durbin Algorithm

Since for large sample sizes, the Cholesky algorithm could be inefficient, faster methods for calculating the log-likelihood function have been developed. These numerical procedures, designed to exploit the Toeplitz structure of the variance–covariance matrix of an second-order stationary process, are based on the seminal works by Levinson (1947) and Durbin (1960).

Let $\hat{y}_1 = 0$ and $\hat{y}_{t+1} = \phi_{t1} y_t + \cdots + \phi_{tt} y_1$ for $t = 1, \ldots, n-1$ be the one-step predictors of the process $\{y_t\}$ based on the finite past $\{y_1, \ldots y_{t-1}\}$, where the partial regression coefficients ϕ_{tj} are given by the recursive equations

$$\begin{aligned}
&\phi_{tt} = \left[\gamma(t) - \sum_{i=1}^{t-1} \phi_{t-1,i}\gamma(t-i)\right] / v_{t-1} \\
&\phi_{tj} = \phi_{t-1,j} - \phi_{tt}\phi_{t-1,t-j}, \qquad j = 1, \ldots, n-1 \\
&v_0 = \gamma(0) \\
&v_t = v_{t-1}[1 - \phi_{tt}^2], \qquad t = 1, \ldots, n-1
\end{aligned}$$

Now, if $e_t = y_t - \hat{y}_t$ is the prediction error and $\boldsymbol{e} = (e_1, \ldots, e_n)'$, then $\boldsymbol{e} = L_2 Y$, where L_2 is the lower triangular matrix

$$L_2 = \begin{pmatrix} 1 & & & & \\ -\phi_{11} & 1 & & & \\ -\phi_{22} & -\phi_{21} & 1 & & \\ -\phi_{33} & -\phi_{32} & -\phi_{31} & 1 & \\ \vdots & \vdots & & & \\ -\phi_{n-1,n-1} & -\phi_{n-1,n-2} & \cdots & -\phi_{n-1,1} & 1 \end{pmatrix}$$

Thus, Γ_θ may be decomposed as $\Gamma_\theta = L_2 D L_2'$, where $D = \text{diag}(v_0, \ldots, v_{n-1})$. Consequently, $\det \Gamma_\theta = \prod_{j=1}^{n} v_{j-1}$ and $Y'\Gamma_\theta^{-1} Y = \boldsymbol{e}' D^{-1} \boldsymbol{e}$. Hence, the log-likelihood function may be written as

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{n} \log v_{t-1} - \frac{1}{2}\sum_{t=1}^{n} e_t^2 / v_{t-1}$$

The arithmetic complexity of this algorithm is $\mathcal{O}(2n^2)$ for a linear stationary process, see Ammar (1996). However, for some Markovian processes such as the family of ARMA models, the Levinson–Durbin algorithm can be implemented in only $\mathcal{O}(n)$ operations, see for example Section 5.3 of Brockwell and Davis (1991). Unfortunately, ARFIMA models are not Markovian and this reduction in operations count does not apply to them.

2.3. Calculation of Autocovariances

A critical aspect in the implementation of the Cholesky and the Levinson–Durbin algorithms is the calculation of the ACF process. A closed form expression for the ACF of an ARFIMA model is given in Sowell (1992). Here, we briefly review this and other alternative methods of obtaining the ACF of a long-memory process.

Observe that the polynomial $\Phi(B)$ defined in (1) can be written as

$$\Phi(B) = \prod_{i=1}^{p} (1 - \rho_i B)$$

where $\{\rho_i\}$ are the roots of the polynomial $\Phi(z^{-1})$. Assuming that all these roots have multiplicity one, it can be deduced from (2) that

$$\gamma(k) = \sigma_\varepsilon^2 \sum_{i=-q}^{q} \sum_{j=1}^{p} \psi(i)\xi_j C(d, p + i - k,\ \rho_j)$$

with $\psi(i) = \sum_{j=\max(0,l)}^{\min(q,q-1)} \theta_j \theta_{j-l}$, $\xi_j = \left[\rho_j \prod_{i=1}^{p}(1-\rho_i\rho_j)\prod_{m\neq j}(\rho_j-\rho_m)\right]^{-1}$ and

$$c(d,h,\rho) = \frac{\Gamma(1-2d)\Gamma(d+h)}{\Gamma(1-d+h)\Gamma(1-d)\Gamma(d)} \times [\rho^{2p}F(d+h,\ 1,\ 1-d+h,\ \rho) + F(d-h,\ 1,\ 1-d-h,\ \rho) - 1]$$

where $\Gamma(\cdot)$ is the Gamma function and $F(a, b, c, x)$ is the Gaussian hypergeometric function, see (Gradshteyn & Ryzhik, 2000, Section 9.1).

An alternative method for calculating the ACF is the so-called splitting method, see Bertelli and Caporin (2002). This technique is based on the decomposition of the model into its ARMA and its fractional integrated (FI) parts. Let $\gamma_1(\cdot)$ be the ACF of the ARMA component and $\gamma_2(\cdot)$ be the ACF of the fractional noise. Then, the ACF of the corresponding ARFIMA process is given by the convolution of these two functions:

$$\gamma(k) = \sum_{h=-\infty}^{\infty} \gamma_1(h)\gamma_2(k-h)$$

If this infinite sum is truncated to m summands, we obtain the approximation

$$\gamma(k) \approx \sum_{h=-m}^{m} \gamma_1(h)\gamma_2(k-h)$$

Thus, the ACF can be efficiently calculated with a great level of precision, see for instance the numerical experiments reported by Bertelli and Caporin (2002) for further details.

2.4. Exact State-Space Method

Another method for computing exact ML estimates is provided by the state-space system theory. In this Section, we review the application of Kalman filter techniques to long-memory processes. Note that since these processes are not Markovian, all the state space representations are infinite dimensional as shown by Chan and Palma (1998). Despite this fact, the Kalman filter equations can be used to calculate the exact log-likelihood (3) in a finite number of steps.

Recall that a causal ARFIMA(p, d, q) process $\{y_t\}$ has a linear process representation given by

$$y_t = \frac{\Theta(B)}{\Phi(B)}(1-B)^{-d}\varepsilon_t = \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j} \tag{4}$$

where φ_j are the coefficients of $\varphi(z) = \sum_{j=0}^{\infty} \varphi_j z^j = \Theta(z)\Phi(z)^{-1}(1-z)^{-d}$. An infinite-dimensional state-space representation may be constructed as follows. From Eq. (4), a state space system may be written as

$$X_{t+1} = FX_t + H\varepsilon_t \tag{5}$$

$$y_t = GX_t + \varepsilon_t \tag{6}$$

where

$$X_t = [y(t|t-1)\ y(t+1|t-1)\ y(t+2|t-1)\ \ldots\ ]' \tag{7}$$

$$y(t|j) = E[y_t|y_j, y_{j-1}, \ldots] \tag{8}$$

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \ddots & & \end{bmatrix}, \qquad H = [\varphi_1, \varphi_2, \ldots]', \text{ and } G = [1, 0, 0, \ldots] \tag{9}$$

The log-likelihood function can be evaluated by directly applying the Kalman recursive equations in Proposition 12.2.2 of Brockwell and Davis (1991) to the infinite-dimensional system. The Kalman algorithm is as follows: Let $\Omega_t = (\omega_{ij}^{(t)})$ be the state estimation error covariance matrix at time t. The Kalman equations for the infinite-dimensional system are given by

$$\hat{X}_1 = E[X_1] \tag{10}$$

$$\Omega_1 = E[X_1 X_1'] - E[\hat{X}_1 \hat{X}'_1] \tag{11}$$

$$\Delta_t = \omega_{11}^{(t)} + 1 \tag{12}$$

$$\omega_{ij}^{(t+1)} = \omega_{i+1,j+1}^{(t)} + \varphi_i \varphi_j - \frac{(\omega_{i+1,1}^{(t)} + \varphi_i)(\omega_{j+1,1}^{(t)} + \varphi_j)}{\omega_{11}^{(t)} + 1} \tag{13}$$

and

$$\hat{X}_{t+1} = (\hat{X}_1^{(t+1)}, \hat{X}_2^{(t+1)}, \ldots)' = (\hat{X}_i^{(t+1)})'_{i=1,2,\ldots} \tag{14}$$

where

$$\hat{X}_i^{(t+1)} = \hat{X}_{i+1}^{(t)} + \frac{(y_t - \hat{X}_1^{(t)})(\omega_{i+1,1}^{(t)} + \varphi_i)}{\omega_{11}^{(t)} + 1} \tag{15}$$

$$\hat{y}_{t+1} = G\hat{X}_{t+1} = \hat{X}_1^{(t+1)} \tag{16}$$

and the log-likelihood function is given by

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2}\left\{ n \log 2\pi + \sum_{t=1}^{n} \log \Delta_t + n \log \sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2} \sum_{t=1}^{n} \frac{(y_t - \hat{y}_t)^2}{\Delta_t} \right\}$$

Although the state space representation of an ARFIMA model is infinite-dimensional, the exact likelihood function can be evaluated in a finite number of steps. Specifically, we have the following theorem due to Chan and Palma (1998).

Theorem 1. Let $\{y_1, \ldots, y_n\}$ be a finite sample of an ARFIMA(p,d,q) process. If Ω_1 is the variance of the initial state X_1 of the infinite-dimensional representation (5)–(9), then the computation of the exact likelihood function (3) depends only on the first n components of the Kalman Eqs. (10)–(16).

It is worth noting that as a consequence of Theorem 1, given a sample of n observations from an ARFIMA process, the calculation of the exact likelihood function is based only on the first n components of the state vector. Therefore, the remaining infinitely many components of the state vector can be omitted from the computations.

The arithmetic complexity of this algorithm is $\mathcal{O}(n^3)$. Therefore, it is comparable to the Cholesky decomposition but it is less efficient than the Levinson–Durbin procedure. The Kalman approach is advisable for moderate sample sizes or for handling time series with missing values. The state space framework provides a simple solution to this problem. Note that the Levinson–Durbin method is no longer appropriate when the series displays missing values since the variance–covariance matrix of the incomplete data does not have a Toeplitz structure.

2.5. Asymptotic Results for the Exact MLE

The consistency, asymptotic normality and efficiency of the MLE have been established by Yajima (1985) for the fractional noise process and by Dahlhaus (1989) for the general case including the ARFIMA model.

Let $\hat{\boldsymbol{\theta}}_n$ be the value that maximizes the exact log-likelihood where

$$\boldsymbol{\theta} = (\phi_1, \ldots, \phi_p, \phi_1, \ldots, \theta_q, d, \sigma_\varepsilon)'$$

is a $p + q + 2$ dimensional parameter vector and let $\boldsymbol{\theta}_0$ be the *true* parameter. Assume that the regularity conditions listed in Dahlhaus (1989) hold.

Theorem 2. (Consistency) $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability as $n \to \infty$.
(Central Limit Theorem) $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N(0, \Gamma^{-1}(\boldsymbol{\theta}_0))$, as $n \to \infty$, where $\Gamma(\boldsymbol{\theta}) = (\Gamma_{ij}(\boldsymbol{\theta}))$ with

$$\begin{aligned} &\Gamma_{ij}(\boldsymbol{\theta}) = \tfrac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\frac{\partial \log k(\omega,\boldsymbol{\theta})}{\partial\theta_i}\right\}\left\{\frac{\partial \log k(\omega,\boldsymbol{\theta})}{\partial\theta_j}\right\}dw \\ &\text{and } k(\omega,\boldsymbol{\theta}) = |\textstyle\sum_{j=0}^{\infty}\psi_j(\boldsymbol{\theta})e^{ij\omega}|^2 \end{aligned} \tag{17}$$

(Efficency) $\hat{\boldsymbol{\theta}}_n$ is an efficient estimator of $\boldsymbol{\theta}_0$.

The next two sections discuss ML methods based on AR and MA approximations.

3. AUTOREGRESSIVE APPROXIMATIONS

Since the computation of exact ML estimates is computationally highly demanding, several authors have considered the use of AR approximations to speed up the calculation of parameter estimates. In particular, this approach has been adopted by Granger and Joyeux (1980), Li and McLeod (1986), Haslett and Raftery (1989), Beran (1994b), Shumway and Stoffer (2000) and Bhansali and Kokoszka (2003), among others.

Most of these techniques are based on the following estimation strategy. Consider a long-memory process $\{y_t\}$ defined by the AR(∞) expansion

$$y_t - \pi_1(\boldsymbol{\theta})y_{t-1} - \pi_2(\boldsymbol{\theta})y_{t-2} - \cdots = \varepsilon_t$$

where $\pi_j(\boldsymbol{\theta})$ are the coefficients of $\Phi(B)\Theta^{-1}(B)(1-B)^d$. In practice, only a finite number of observations is available, $\{y_1, \ldots, y_n\}$, therefore the following truncated model is considered

$$y_t - \pi_1(\boldsymbol{\theta})y_{t-1} - \cdots - \pi_m(\boldsymbol{\theta})y_{y-m} = \tilde{\varepsilon}_t$$

for $m < t \le n$. The ML estimate $\hat{\boldsymbol{\theta}}_n$ is found by minimizing

$$\mathscr{L}_0(\boldsymbol{\theta}) = \sum_{t=m+1}^{n}[y_t - \pi_1(\boldsymbol{\theta})y_{t-1} - \cdots - \pi_m(\boldsymbol{\theta})y_{t-m}]^2$$

Upon this basic framework, many refinements can be made to improve the quality of these estimates. In what follows next, we describe some of these refinements. For simplicity, any estimator produced by the maximization of an approximation of the Gaussian likelihood function (3) will be called QMLE.

3.1. Haslett–Raftery Method

The following technique was proposed by Haslett and Raftery (1989). Consider the ARFIMA process (1). An approximate one-step predictor of y_t is given by

$$\hat{y}_t = \Phi(B)\Theta(B)^{-1} \sum_{j=1}^{t-1} \phi_{tj} y_{t-j} \tag{18}$$

with prediction error variance

$$v_t = \mathrm{var}(y_t - \hat{y}_t) = \sigma_y^2 \kappa \prod_{j=1}^{t-1} (1 - \phi_{jj}^2)$$

where $\sigma_y^2 = \mathrm{var}(y_t)$, κ is the ratio of the innovations variance to the variance of the ARMA(p,q) process as given by Eq. (3.4.4) of Box, Jenkins, and Reinsel (1994) and

$$\phi_{tj} = -\binom{t}{j} \frac{\Gamma(j-d)\Gamma(t-d-j+1)}{\Gamma(-d)\Gamma(t-d+1)}, \quad \text{for } j = 1, \ldots, t$$

To avoid the computation of a large number of coefficients ϕ_{tj}, the last term of the predictor (18) is approximated by

$$\sum_{j=1}^{t-1} \phi_{tj} y_{t-j} \approx \sum_{j=1}^{M} \phi_{tj} y_{t-j} - \sum_{j=M+1}^{t-1} \pi_j y_{t-j} \tag{19}$$

since $\phi_{tj} \sim -\pi_j$ for large j, cf. Hosking (1981), where for simplicity π_j denotes $\pi_j(\boldsymbol{\theta})$ and $a_j \sim b_j$ means that $a_j/b_j \to 1$, as $j \to \infty$.

An additional approximation is made to the second term on the right-hand-side of (19):

$$\sum_{j=M+1}^{t-1} \pi_j y_{t-j} \approx M \pi_M d^{-1} \left[1 - \left(\frac{M}{t} \right)^d \right] \bar{y}_{M+1,t-1-M}$$

where $\bar{y}_{M+1,t-1-M} = \frac{1}{t-1-2M} \sum_{j=M+1}^{t-1-M} y_j$. Hence, a QMLE $\hat{\boldsymbol{\theta}}_n$ is obtained by maximizing

$$\mathscr{L}_1(\boldsymbol{\theta}) = \text{constant} - \frac{1}{2} n \log[\hat{\sigma}_\varepsilon^2(\boldsymbol{\theta})]$$

with

$$\hat{\sigma}_\varepsilon^2(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \hat{y}_t)^2}{v_t}$$

The arithmetic complexity of the Haslett and Raftery method is of order $\mathcal{O}(nM)$. For a fixed M, the algorithm is of order $\mathcal{O}(n)$, which is much faster compared to the Levinson–Durbin method. Haslett and Raftery (1989) suggest $M = 100$. Note that, when $M = n$, the exact ML estimated is obtained. However, the arithmetic complexity in that case becomes $\mathcal{O}(n^2)$ and no gain is obtained as compared to the Levinson–Durbin approach.

3.2. Beran Method

Beran (1994a) proposed the following version of the AR approximation approach. Assume that the following Gaussian innovation sequence

$$\varepsilon_t = y_t - \sum_{j=1}^{\infty}\pi_j(\boldsymbol{\theta})y_{t-j}$$

Since the values $\{y_t,\ t \le 0\}$ are not observed, an approximate innovation sequence $\{u_t\}$ can be obtained by assuming that $y_t = 0$ for $t \le 0$,

$$u_t(\boldsymbol{\theta}) = y_t - \sum_{j=1}^{t-1}\pi_j(\boldsymbol{\theta})y_{t-j}$$

for $t = 2, \ldots, n$. Let $r_t(\boldsymbol{\theta}) = u_t(\boldsymbol{\theta})/\theta_1$ and $\boldsymbol{\theta} = (\sigma_\varepsilon,\ \phi_1, \ldots, \phi_p,\ \theta_1, \ldots, \theta_q,\ d)$. Then, a QMLE for $\boldsymbol{\theta}$ is provided by the minimization of

$$\mathcal{L}_2(\boldsymbol{\theta}) = 2n\ \log(\theta_1) + \sum_{t=2}^{n} r_t^2(\boldsymbol{\theta})$$

Now, by taking partial derivatives with respect to $\boldsymbol{\theta}$, the minimization problem is equivalent to solving the non-linear equations

$$\sum_{t=2}^{n}\{r_t(\boldsymbol{\theta})\dot{r}_t(\boldsymbol{\theta}) - E[r_t(\boldsymbol{\theta})\dot{r}_t(\boldsymbol{\theta})]\} = 0 \tag{20}$$

where $\dot{r}_t(\boldsymbol{\theta}) = \left(\frac{\partial r_t(\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial r_t(\boldsymbol{\theta})}{\partial\theta_r}\right)'$.

The arithmetic complexity of this method is $\mathcal{O}(n^2)$, that is, comparable to the Levinson–Durbin algorithm. Unlike the Haslett and Raftery method, the Beran approach uses the same variance for all the errors u_t. Hence, its performance may be poor for short time series.

3.2.1. Asymptotic Behavior

The QMLE based on the AR approximations share the same asymptotic properties with the exact MLE. The following results are due to Beran (1994a). Let $\tilde{\boldsymbol{\theta}}_n$ be the value that solves (20). Then,

Theorem 3. (Consistency) $\tilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$, in probability, as $n \to \infty$.
(Central Limit Theorem) $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N(0, \Gamma^{-1}(\boldsymbol{\theta}_0))$, as $n \to \infty$, with $\Gamma(\boldsymbol{\theta}_0)$ is given in (17).
(Efficiency) $\tilde{\boldsymbol{\theta}}_n$ is an efficient estimator of $\boldsymbol{\theta}_0$.

4. MOVING AVERAGE APPROXIMATIONS

A natural alternative to AR approximations is the truncation of the infinite MA expansion of a long-memory process. The main advantage of this approach is the easy implementation of the Kalman filter recursions and the simplicity of the analysis of the theoretical properties of the ML estimates. Furthermore, if differencing is applied to the series, then the resulting truncation has less error variance than the AR approximation.

A causal representation of an ARFIMA(p, d, q) process $\{y_t\}$ is given by

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \tag{21}$$

On the other hand, we may consider an approximate model for (21) given by

$$y_t = \sum_{j=0}^{m} \psi_j \varepsilon_{t-j} \tag{22}$$

which corresponds to a MA(m) process in contrast to the MA(∞) process (21). A canonical state space representation of the MA(m) model (22) is given by

$$X_{t+1} = FX_t + H\varepsilon_t$$
$$y_t = GX_t + \varepsilon_t$$

with

$$F = \begin{bmatrix} 0 & I_{m-1} \\ 0 \ldots & 0 \end{bmatrix}, \qquad G = [1\ 0\ 0\ \ldots\ 0], \qquad H = [\psi_1\ \ldots\ \psi_m]'$$

$$X_t = [y(t|t-1),\ y(t+1|t-1),\ \ldots,\ y(t+m-1|t-1)]'$$
$$y(t+j|t-1) = E[y_{t+j}|y_{t-1}, y_{t-2}, \ldots]$$

The approximate representation of a causal ARFIMA(p, d, q) has computational advantages over the exact one. In particular, the order of the MLE algorithm is reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$. A brief discussion about the Kalman filter implementation of this state space system follows.

4.1. Kalman Recursions

Let the initial conditions be $\hat{X}_1 = E[X_1]$ and $\Omega_1 = E[X_1 X_1'] - E[\hat{X}_1 \hat{X}'_1]$. The recursive Kalman equations may be written as follows (cf. Chan, 2002, Section 11.3):

$$\Delta_t = G\Omega_t G' + \sigma_\varepsilon^2 \tag{23}$$

$$\Theta_t = F\Omega_t G' + S \tag{24}$$

$$\Omega_{t+1} = F\Omega_t F' + Q - \Theta_t \Delta_t^{-1} \Theta_t' \tag{25}$$

$$\begin{aligned} \hat{X}_{t+1} &= F\hat{X}_t + \Theta_t \Delta_t^{-1}(y_t - G\hat{X}_t) \\ \hat{y}_t &= G\hat{X}_t \end{aligned} \tag{26}$$

for $t = 1, 2, \ldots, n$, where $Q = \mathrm{var}(H\varepsilon_t)$, $\sigma_\varepsilon = \mathrm{var}(\varepsilon_t)$ and $S = \mathrm{cov}(H\varepsilon_t, \varepsilon_t)$. The log-likelihood function, excepting a constant, is given by

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\left\{ \sum_{t=1}^{n} \log\, \Delta_t(\theta) + \sum_{t=1}^{n} \frac{(y_t - \hat{y}_t(\theta))^2}{\Delta_t(\theta)} \right\}$$

where $\boldsymbol{\theta} = (\phi_1, \ldots \phi_p, \theta_1, \ldots, \theta_q, d, \sigma^2)$ is the parameter vector associated with the ARFIMA representation (1). In order to evaluate the log-likelihood function $\mathcal{L}(\boldsymbol{\theta})$, we may choose the initial conditions $\hat{X}_1 = E[X_1] = 0$ and $\Omega_1 = E[X_1\, X_1'] = [\omega(i, j)]_{i,j=1,2,\ldots,}$ where $\omega(i, j) = \sum_{k=0}^{\infty} \psi_{i+k}\psi_{j+k}$.

The evolution of the state estimation and its variance, Ω_t, is given by the following recursive equations. Let $\delta_i = 1$ if $i \in \{0, 1, \ldots, m-1\}$ and $\delta_i = 0$ otherwise. Furthermore, let $\delta_{ij} = \delta_i \delta_j$. Then, the elements of Ω_{t+1} and $\hat{X}_{t+1}$ in (25) and (26) are as follows:

$$\begin{aligned} \omega_{t+1}(i, j) = {} & \omega_t(i+1, j+1)\delta_{ij} + \psi_i \psi_j \\ & - \frac{[\omega_t(i+1, 1)\delta_i + \psi_i][\omega_t(j+1, 1)\delta_j + \psi_j]}{\omega_t(1, 1) + 1} \end{aligned}$$

and

$$\hat{X}_{t+1}(i) = \hat{X}_t(i+1)\delta_i + \frac{(\omega_t(i+1,1)\delta_i + \psi_i)(y_t - \hat{X}_t(1))}{\omega_t(1,1)+1}$$

In addition,

$$\hat{y}_t = G\hat{X}_t = \hat{X}_t(1)$$

In order to speed up the algorithm, we can difference the series $\{y_t\}$ so that the MA expansion of the differenced series converges faster. To this end, consider the MA expansion of the differenced process

$$z_t = (1-B)y_t = \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j} \tag{27}$$

where $\varphi_j = \psi_j - \psi_{j-1}$. If we truncate this expansion after m components, an approximate model can be written as

$$z_t = \sum_{j=0}^{m} \varphi_j \varepsilon_{t-j} \tag{28}$$

The main advantage of this approach is that the coefficients φ_j converge to zero faster than the coefficients ψ_j do. Thus, a smaller truncation parameter is necessary to achieve a good level of approximation.

A state space representation of this truncated model is given by, see for example Section 12.1 of Brockwell and Davis (1991),

$$X_{t+1} = \begin{bmatrix} 0 & I_{m-1} \\ 0 \cdots & 0 \end{bmatrix} X_t + \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_m \end{bmatrix} \varepsilon_t$$

and

$$z_t = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix} X_t + \varepsilon_t$$

The Gaussian log-likelihood of the truncated model (28) may be written as

$$\mathscr{L}_n(\boldsymbol{\theta}) = \frac{1}{2n} \log \det T_{n,m}(\boldsymbol{\theta}) - \frac{1}{2n} Z_n' T_{n,m}(\boldsymbol{\theta})^{-1} Z_n$$

where $[T_{n,m}(\boldsymbol{\theta})]_{r,s=1,\dots,n} = \int_{-\pi}^{\pi} \tilde{f}_{m,\boldsymbol{\theta}}(\omega) e^{i\omega(r-s)} d\omega$ is the covariance matrix of $Z_n = (z_1 \dots z_n)'$ with $\tilde{f}_{m,\boldsymbol{\theta}}(\omega) = \sigma_\varepsilon^2 |\varphi_m(e^{i\omega})|^2$ and $\varphi_m(e^{i\omega}) = 1 + \varphi_1 e^{i\omega} + \cdots + \varphi_{1_m} e^{mi\omega}$.

In this case, the matrices involved in the truncated Kalman equations are of order $m \times m$. Therefore, only m^2 evaluations are required for each iteration and the algorithm has an order $n \times m^2$. For a fixed truncation parameter m, the calculation of the likelihood function is only of order $\mathcal{O}(n)$ for the approximate ML method. Thus, for very large samples, it may be desirable to consider truncating the Kalman recursive equations after m components. With this truncation, the number of operations required for a single evaluation of the log-likelihood function is reduced to an order of $\mathcal{O}(n)$.

This approach is discussed in Chan and Palma (1998), where they show the following result.

Theorem 4. (Consistency) Assume that $m = n^\beta$ with $\beta > 0$, then as $n \to \infty$, $\tilde{\boldsymbol{\theta}}_{n,m} \to \boldsymbol{\theta}_0$, in probability.
(Central Limit Theorem) Suppose that $m = n^\beta$ with $\beta \geq 1/2$, then as $n \to \infty$, $\sqrt{n}(\tilde{\boldsymbol{\theta}}_{n,m} - \boldsymbol{\theta}_0) \to N(0, \Gamma^{-1}(\boldsymbol{\theta}_0))$,
where $\Gamma(\boldsymbol{\theta})$ is given in (17).
(Efficiency) Assume that $m = n^\beta$ with $\beta \geq 1/2$, then $\tilde{\boldsymbol{\theta}}_{n,m}$, is an efficient estimator of θ_0.

Observe that both AR(m) and MA(m) approximations produce algorithms with arithmetic complexity of order $\mathcal{O}(n)$. However, the quality of the approximation is governed by the truncation parameter m. Bondon and Palma (2005) prove that the variance of the truncation error for an AR(m) approximation is of order $\mathcal{O}(1/n)$. On the other hand, it can be easily shown that for the MA(m) case, this quantity is of order $\mathcal{O}(n^{2d-1})$. Furthermore, when the differenced approach is used, the truncation error variance is of order $\mathcal{O}(n^{2d-3})$.

5. WHITTLE APPROXIMATIONS

Another approach to obtain approximate ML estimates is based on the calculation of the periodogram by means of the Fast Fourier Transform (FFT) and the use of the Whittle approximation of the Gaussian log-likelihood function. This approach produces fast numerical algorithms for computing parameter estimates, since the calculation of the FFT has an arithmetic complexity $\mathcal{O}(n \log_2(n))$ (cf. Press et al., 1992, p. 498).

5.1. Whittle Approximation of the Gaussian Likelihood Function

Consider the Gaussian process $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ with zero mean and variance $\Gamma_{\boldsymbol{\theta}}$. Then the log-likelihood function divided by the sample size is given by

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2n}\log\ \det\ \Gamma_\theta - \frac{1}{2n}\, Y'\Gamma_\theta^{-1}\, Y \tag{29}$$

Observe that the variance–covariance matrix $\Gamma_{\boldsymbol{\theta}}$ can be expressed in terms of the spectral density of the process $f_{\boldsymbol{\theta}}(\cdot)$ as follows:

$$(\Gamma_{\boldsymbol{\theta}})_{ij} = \gamma_{\boldsymbol{\theta}}(i-j)$$

where

$$\gamma_{\boldsymbol{\theta}}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{\boldsymbol{\theta}}(\omega)\ \exp\,(i\omega k)\, d\omega$$

In order to obtain the method proposed by Whittle (1951), two approximations are made. Following the result by Grenander and Szegö (1958) that

$$\frac{1}{n}\log\ \det\ \Gamma_{\boldsymbol{\theta}} \to \frac{1}{2\pi}\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\omega)\, d\omega$$

as $n \to \infty$, the first term in (29) is approximated by

$$\frac{1}{2n}\log\ \det\ \Gamma_\theta \approx \frac{1}{4\pi}\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\omega)\, d\omega$$

On the other hand, the second term in (29) is approximated by

$$\begin{aligned}
\frac{1}{2\pi}\, Y'\Gamma_\theta^{-1}\, Y &\approx \sum_{l=1}^{n}\sum_{j=1}^{n} y_l \left\{\frac{1}{4\pi n}\int_{-\pi}^{\pi} f_{\boldsymbol{\theta}}^{-1}(\omega)\ \exp\,[i\omega(l-j)]d\omega\right\} y_j \\
&= \frac{1}{4\pi n}\int_{-\pi}^{\pi} f_{\boldsymbol{\theta}}^{-1}(\omega) \sum_{l=1}^{n}\sum_{j=1}^{n} y_l y_j\ \exp\,[i\omega(l-j)]d\omega \\
&= \frac{1}{4\pi n}\int_{-\pi}^{\pi} f_{\boldsymbol{\theta}}^{-1}(\omega) \Big|\sum_{j=1}^{n} y_j\ \exp\,(i\omega j)\Big|^2 d\omega \\
&= \frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{I(\omega)}{f_{\boldsymbol{\theta}}(\omega)} d\omega
\end{aligned}$$

where $I(\omega) = \frac{1}{n}\left|\sum_{j=1}^{n}(y_j - \bar{y})\ \exp\,(i\omega j)\right|^2$ is the periodogram of the series $\{y_t\}$. Thus, the log-likelihood function is approximated by

$$\mathscr{L}_3(\boldsymbol{\theta}) = -\frac{1}{4\pi}\left\{\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\omega) d\omega + \int_{-\pi}^{\pi} \frac{I(\omega)}{f_{\boldsymbol{\theta}}(\omega)} d\omega\right\} \tag{30}$$

5.2. Discrete Version

The evaluation of the log-likelihood function (30) requires the calculation of integrals. To simplify this computation, the integrals can be substituted by Riemann sums as follows:

$$\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\omega) d\omega \approx \frac{2\pi}{n} \sum_{j=1}^{n} \log f_{\boldsymbol{\theta}}(\omega_j)$$

and

$$\int_{-\pi}^{\pi} \frac{I(\omega)}{f_{\boldsymbol{\theta}}(\omega)} d\omega\} \approx \frac{2\pi}{n} \sum_{j=1}^{n} \frac{I(\omega_j)}{f_{\boldsymbol{\theta}}(\omega_j)}$$

where $\omega_j = \frac{2\pi j}{n}$ are the Fourier frequencies. Thus, a discrete version of the log-likelihood function (30) is

$$\mathscr{L}_4(\boldsymbol{\theta}) = -\frac{1}{2n} \left\{ \sum_{j=1}^{n} \log f_{\boldsymbol{\theta}}(\omega_j) + \sum_{j=1}^{n} \frac{I(\omega_j)}{f_{\boldsymbol{\theta}}(\omega_j)} \right\}$$

5.3. Alternative Versions

Further simplifications of the Whittle log-likelihood function can be made. For example, by assuming that the spectral density is normalized as

$$\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\omega) d\omega = 0 \tag{31}$$

then the Whittle log-likelihood function is reduced to

$$\mathscr{L}_5(\boldsymbol{\theta}) = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{I(\omega)}{f_{\boldsymbol{\theta}}(\omega)} d\omega$$

with the discrete version

$$\mathscr{L}_6(\boldsymbol{\theta}) = -\frac{1}{2n} \sum_{j=1}^{n} \frac{I(\omega_j)}{f_{\boldsymbol{\theta}}(\omega_j)}$$

Observe that by virtue of the well-known Szegö–Kolmogorov formula

$$\sigma_\varepsilon^2 = 2\pi \exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\omega) d\omega \right]$$

the normalization (31) is equivalent to setting $\sigma_\varepsilon^2 = 2\pi$.

5.4. Asymptotic Results

The asymptotic behavior of the Whittle estimator is similar to that of the exact MLE. The following theorem combines results by Fox and Taqqu (1986) and Dahlhaus (1989).

Theorem 5. Let $\hat{\boldsymbol{\theta}}_n^{(i)}$ be the value that maximizes the log-likelihood function $\mathscr{L}_i(\boldsymbol{\theta})$ for $i = 3, \ldots, 6$ for a Gaussian process $\{y_t\}$. Then, under some regularity conditions, $\hat{\boldsymbol{\theta}}_n^{(i)}$ is consistent and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0) \to N(0, \Gamma_{\boldsymbol{\theta}_0}^{-1})$, as n $\to \infty$.

5.5. Non-Gaussian Processes

All of the above methods apply to Gaussian processes. When this assumption in dropped, it is still possible to find well-behaved Whittle estimates. In particular, Giraitis and Surgailis (1990) have studied the estimates based on the maximization of the log-likelihood function $\mathscr{L}_5(\boldsymbol{\theta})$ for a general class of linear processes with independent innovations.

Consider the process $\{y_t\}$ generated by the Wold decomposition

$$y_t = \sum_{j=0}^{\infty} \psi_j(\boldsymbol{\theta}) \varepsilon_{t-j}$$

where ε_t is an i.i.d. sequence with finite four cumulant and $\sum_{j=0}^{\infty} \psi_j^2(\boldsymbol{\theta}) < \infty$. The following result, due to Giraitis and Surgailis (1990), establishes the consistency and the asymptotic normality of the Whittle estimate under these circumstances.

Theorem 6. Let $\hat{\boldsymbol{\theta}}_n$ be the value that maximizes the log-likelihood function $\mathscr{L}_5(\boldsymbol{\theta})$. Then, under some regularity conditions, $\hat{\boldsymbol{\theta}}_n$ is consistent and $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \to N(0, \Gamma_{\boldsymbol{\theta}_0}^{-1})$, as $n \to \infty$.
Note that this theorem does not assume the normality of the process.

5.6. Semi-Parametric Methods

Another generalization of the Whittle estimator is the Gaussian semi-parametric estimation method proposed by Robinson (1995). This estimation approach does not require the specification of a parametric model for the data. It only relies on the specification of the shape of the spectral density of the time series.

Consider the stationary process $\{y_t\}$ with spectral density satisfying

$$f(\omega) \sim G\omega^{1-2H}$$

as $\omega \to 0+$, with $G \in (0, \infty)$ and $H \in (0, 1)$. Note that for an ARFIMA model, the terms G and H correspond to $\sigma^2/2\pi[\theta(1)/\phi(1)]^2$ and $1/2 + \mathrm{d}$, respectively. H is usually called the *Hurst parameter.*

Let the objective function $Q(G, H)$ to be minimized be given by

$$Q(G, H) = \frac{1}{m}\sum_{j=1}^{m}\left\{\log\ G\omega_j^{1-2H} + \frac{\omega_j^{2H-1}}{G} I(\omega_j)\right\}$$

where m is an integer satisfying $m < n/2$. Let $(\hat{G}, \hat{H})$ be the value that minimizes $Q(G, H)$. Then, under some regularity conditions of the spectral density and

$$\frac{1}{m} + \frac{m}{n} \to 0$$

as $n \to \infty$, the following result due to Robinson (1995) holds:

Theorem 7. Let H_0 be the true value of the Hurst parameter. The estimator $\hat{H}$ is consistent and $\sqrt{m}(\hat{H} - \hat{H}_0) \to N\left(0, \frac{1}{4}\right)$, as $n \to \infty$.

5.7. Numerical experiments

Table 1 displays the results from several simulations comparing five ML estimation methods for Gaussian processes: Exact MLE, Haslett and Raftery's approach, AR(40) approximation, MA(40) approximation and the Whittle method. The process considered is a fractional noise ARFIMA(0, d, 0) with three values of the long-memory parameter:

Table 1. Finite Sample Behavior of ML Estimates of ARFIMA(0, d, 0) Models. Sample Size $n = 250$ and Truncation $m = 40$ for AR(m) and MA(m) Approximations.

d		Exact	HR	AR	MA	Whittle
0.40	mean	0.371210	0.372320	0.376730	0.371310	0.391760
	stdev	0.047959	0.048421	0.057392	0.050048	0.057801
0.25	mean	0.229700	0.230400	0.229540	0.229810	0.221760
	stdev	0.051899	0.051959	0.056487	0.051475	0.060388
0.10	mean	0.082900	0.083240	0.083910	0.084410	0.065234
	stdev	0.049260	0.049440	0.052010	0.049020	0.051646

$d = 0.1, 0.25, 0.4$, Gaussian innovations with mean zero and variance 1 and sample size $n = 250$. The mean and standard deviations of the estimates are based on 1,000 repetitions. All the simulations were carried out by means of Splus programs, available upon request.

From this table, it seems that all estimates are somewhat downward biased for the three values of d considered. All the estimators, excepting the Whittle, seem to behave similarly in terms of both bias and standard error. On the other hand, the Whittle method has less bias for $d = 0.4$, but greater bias for $d = 0.1$. Besides, this procedure seems to have greater standard error than the other estimators, for the three values of d. The empirical parameter estimate standard deviations of all the method considered are close to its theoretical value 0.04931.

In the next section we discuss the application of the ML estimation methodology to time series with missing values.

6. ESTIMATION OF INCOMPLETE SERIES

The Kalman filter recursive Eqs. (23)–(26) can be modified to calculate the log-likelihood function for incomplete series, as described in Palma and Chan (1997). In this case, we have

$$\Delta_t = G\Omega_t G' + \sigma_w^2$$
$$\Theta_t = F\Omega_t G' + S$$

$$\Omega_{t+1} = \begin{cases} F\Omega_t F' + Q - \Theta_t \Delta_t^{-1} \Theta_t' & y_t \text{ known} \\ F\Omega_t F' + Q & y_t \text{ missing} \end{cases}$$

$$\hat{X}_{t+1} = \begin{cases} F\hat{X}_t + \Theta_t \Delta_t^{-1}(y_t - G\hat{X}_t) & y_t \text{ known} \\ F\hat{X}_t & y_t \text{ missing} \end{cases}$$

Let K_n be the set indexing the observed values of the process $\{y_t\}$. The Kalman recursive log-likelihood function is given by

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2}\left\{ r \log 2\pi + \sum_{t \in K_n} \log \Delta_t + r \log \sigma_\varepsilon^2 \frac{1}{\sigma_\varepsilon^2} \sum_{t \in K_n} \frac{(y_t - \hat{y}_t)^2}{\Delta_t} \right\}$$

where r is the number of observed values and $\Delta_t = \omega_{11}^{(t)} + 1$ is the variance of the best predictor $\hat{y}_t$. This form of the likelihood function may be used to efficiently calculate Gaussian ML estimates. One-step predictions and their corresponding standard deviations are obtained directly from the recursive

Kalman filter equations without further computation, see Palma and Chan (1997) for details.

6.1. Effect of Data Irregularities and Missing Values on ML Estimates

To illustrate the potential dramatic effects of data irregularities such as repeated or missing values on the parameter estimation of a long-memory process, consider the well-known Nile river data, cf. Beran (1994b), shown in Fig. 1. Panel (a) displays the original data. From this plot, it seems that in several periods, the data were repeated year after year in order to complete the series. Panel (b) shows the same series, but without those repeated values (filtered series). This procedure has been discussed in Palma and Del Pino (1999).

Table 2 shows the fitted parameters of an ARFIMA(0, d, 0) using an AR(40) approximation along the Kalman filter, for both the original and the data without repetitions.

Observe that for the original data, the estimate of the long-memory parameter is 0.5, indicating that the model has reached the non-stationary boundary. On the other hand, for the filtered data, the estimate of d belongs to the stationary region. Thus, in this particular case, the presence of data

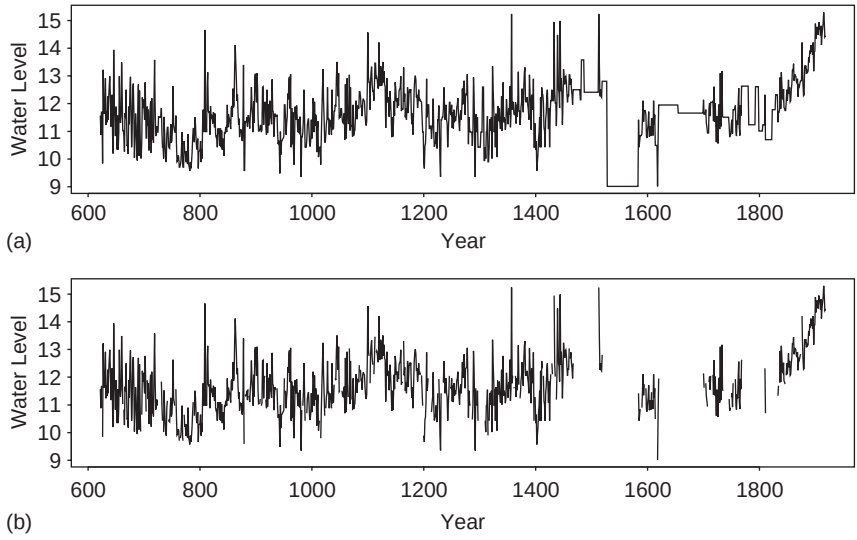


Fig. 1. Nile River Data (622 A.D. – 1921 A.D.). Panel (a) Original Data and panel (b) Filtered Data.

Table 2. Nile River Data: ML Estimates for the Fitted ARFIMA(0, *d,* 0) Model.

Series	$\hat{d}$	$t_{\hat{d}}$	$\hat{\sigma}_\varepsilon$
Panel (a)	0.5000	25.7518	0.6506
Panel (b)	0.4337	19.4150	0.7179

Table 3. Finite Sample Behavior of ML Estimates of ARFIMA(0, *d,* 0) Processes with Missing Values. Sample Size $n = 250$ and Truncation $m = 40$ for AR(m) and MA(m) Approximations.

d	NA's		AR	MA	Expected stdv
	0%	mean	0.380470	0.374390	
		stdev	0.056912	0.048912	0.049312
0.40	15%	mean	0.376730	0.369120	
		stdev	0.060268	0.055014	0.053550
	30%	mean	0.366650	0.363860	
		stdev	0.065166	0.059674	0.058935
	0%	mean	0.225600	0.226900	
		stdev	0.056810	0.052110	0.049312
0.25	15%	mean	0.223500	0.224800	
		stdev	0.065440	0.060040	0.053550
	30%	mean	0.219800	0.220800	
		stdev	0.072920	0.062510	0.058935
	0%	mean	0.081145	0.081806	
		stdev	0.054175	0.051013	0.049312
0.10	15%	mean	0.086788	0.081951	
		stdev	0.058106	0.052862	0.053550
	30%	mean	0.080430	0.081156	
		stdev	0.067788	0.061118	0.058935

irregularities such as the replacement of missing data with repeated values induces non-stationarity. On the other hand, when the missing value is appropriately taken care of, the resulting model is stationary (cf. Palma & Del Pino, 1999).

Table 3 displays the results from Monte Carlo simulations of approximate ML estimates for fractional noise ARFIMA(0, *d,* 0) with missing values at

random. The sample size chosen was $n = 250$ and the AR and MA truncations are $m = 40$ for both cases. The long-memory parameters are $d = 0.1, 0.25, 0.40$ and $\sigma_\varepsilon^2 = 1$. The number of missing values are 38 (15% of the sample) and 75 (30% of the sample) and were selected randomly for each sample.

Note that the bias and the standard deviation of the estimates seem to increase as the number of missing values increases. On the other hand, the sample standard deviation of the estimates seems to be close to the expected values, for the MA approximation. For the AR approximation, these values are greater than expected. The expected standard deviation used here is $\sqrt{6/\pi^2 n^*}$, where n^* is the number of observed values.

7. ESTIMATION OF SEASONAL LONG-MEMORY MODELS

In practical applications, many researchers have found time series exhibiting both long-range dependence and cyclical behavior. For instance, this phenomenon occurs for the inflation rates studied by Hassler and Wolters (1995), revenue series analyzed by Ray (1993), monetary aggregates considered by Porter-Hudak (1990), quarterly gross national product and shipping data discussed by Ooms (1995) and monthly flows of the Nile River studied by Montanari, Rosso, and Taqqu (2000).

Several statistical methodologies have been proposed to model this type of data. For instance, Gray, Zhang, and Woodward (1989) propose the generalized fractional or Gegenbauer (GARMA) processes, Porter-Hudak (1990) discusses seasonal fractionally integrated autoregressive moving average (SARFIMA) models, Hassler (1994) introduces the flexible seasonal fractionally integrated processes (flexible ARFISMA) and Woodward, Cheng, and Gray (1998) introduce the k-GARMA processes. Furthermore, the statistical properties of these models have been investigated by Giraitis and Leipus (1995), Chung (1996), Arteche and Robinson (2000), Velasco and Robinson (2000) and Giraitis, Hidalgo, and Robinson (2001), among others.

A rather general class of Gaussian seasonal long-memory processes is specified by the spectral density

$$f(\omega) = g(\omega)|\omega|^{-\alpha} \prod_{i=1}^{r} \prod_{j=1}^{m_i} |\omega - \omega_{ij}|^{-\alpha_i} \tag{32}$$

where $\omega \in (-\pi, \pi], 0 \leq \alpha$, $\alpha_i < 1$, $i = 1, \ldots, r$, $g(\omega)$ is a symmetric, strictly positive, continuous and bounded function and $\omega_{ij} \neq 0$ are known poles for $j = 1, \ldots, m_i$, $i = 1, \ldots, r$. To ensure the symmetry of f, it is assumed that for any $i = 1, \ldots, r$, $j = 1, \ldots m_i$, there is one and only one $1 \leq j' \leq m_i$ such that $\omega_{ij} = \omega_{ij'}$. The spectral density of many widely used models such as SARFIMA and k-factor GARMA satisfy specification (32).

The exact ML estimation of processes satisfying (32) has been recently studied by Palma and Chan (2005) who have established the following result:

Theorem 8. Let $\hat{\boldsymbol{\theta}}_n$ be the exact MLE for a process satisfying (32) and $\boldsymbol{\theta}_0$ the true parameter. Then, under some regularity conditions we have (Consistency) $\hat{\boldsymbol{\theta}}_{n \to p} \boldsymbol{\theta}_0$ *as* $n \to \infty$. (Central limit theorem) The ML estimate, $\hat{\boldsymbol{\theta}}_n$, satisfies the following limiting distribution as $n \to \infty$: $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N(0, \Gamma(\boldsymbol{\theta}_0)^{-1})$, where $\Gamma(\boldsymbol{\theta}_0)$ is given by (17).
(Efficiency) The ML estimate, $\hat{\boldsymbol{\theta}}_n$, is asymptotically an efficient estimate of $\boldsymbol{\theta}_0$.

7.1. Monte Carlo Studies

In order to assess the finite sample performance of the ML estimates in the context of long-memory seasonal series, a number of Monte Carlo simulations were conducted for the class of SARFIMA$(p, d, q) \times (P, d_s, Q)_s$ models described by the following difference equation (cf. Porter-Hudak, 1990):

$$\phi(B)\Phi(B^s)(1 - B)^d y_t = \theta(B)\Theta(B^s)\varepsilon_t$$

where $\{\varepsilon_t\}$ are standard normal random variables, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, $\Phi(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP}$, $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$, $\Theta(B^s) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{sQ}$, polynomials $\phi(B)$ and $\theta(B)$ have no common zeros, $\Phi(B^s)$ and $\Theta(B^s)$ have no common zeros, and the roots of these polynomials are outside the unit circle.

Table 4 shows the results from simulations for SARFIMA$(0, d, 0) \times (0, d_s, 0)_s$ models. The estimates of $\hat{d}$ and $\hat{d}_s$ reported in columns seven and eight of Table 4 and the estimated standard deviations displayed in the last two columns of the table are based on 1,000 repetitions. The MLE are computed by means of an extension to SARFIMA models of the state space representations of long-memory processes, see Chan and Palma (1998) for details. The theoretical values of the standard deviations of the estimated parameters are based on the formula (17). In general, analytic expressions for the

Table 4. Finite Sample Performance of ML Estimates of SARFIMA(0, d, 0) $\times$ (0, d_s, 0)$_s$ Models for Several Values of s, n, d and d_s.

Period	n	d	d_s	$\sigma(d)$	$\sigma(d_s)$	$\hat{d}$	$\hat{d}_s$	$\hat{\sigma}(\hat{d})$	$\hat{\sigma}(\hat{d}_s)$
4	1000	0.100	0.300	0.025	0.025	0.090	0.299	0.027	0.028
6	500	—	0.300	—	0.035	—	0.286	—	0.036
6	1000	0.150	0.200	0.025	0.025	0.140	0.198	0.026	0.026
6	2000	0.200	0.200	0.018	0.018	0.195	0.196	0.018	0.018
12	3000	—	0.250	—	0.014	—	0.252	—	0.016
12	1000	0.200	0.100	0.025	0.025	0.194	0.094	0.025	0.027

integral in (17) are difficult to obtain for an arbitrary period s. For an SARFIMA(0, d, 0) $\times$ (0, d_s, 0)$_s$ model, the matrix $\Gamma(\boldsymbol{\theta})$ can be written as

$$\Gamma(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\pi^2}{6} & c \\ c & \frac{\pi^2}{6} \end{pmatrix}$$

with $c = \frac{1}{\pi}\int_{-\pi}^{\pi}\{\log|2\sin(\frac{\omega}{2})|\}\{\log|2\sin(s\frac{\omega}{2})|\}d\omega$. An interesting feature of the asymptotic variance of the parameters is that for an SARFIMA(0, d, 0) $\times$ (0, d_s, 0)$_s$ process, the variance of $\hat{d}$ is the same as the variance of $\hat{d}_s$.

From Table 4, note that the estimates and their standard deviations are close to the theoretical values, for all the sample sizes and combinations of parameters investigated.

8. HETEROSKEDASTIC TIME SERIES

Evidence of long-memory behavior in returns and/or empirical volatilities has been observed by several authors, see for example Robinson (1991) and references therein. Accordingly, several models have been proposed in the econometric literature to explain the combined presence of long-range dependence and conditional heteroskedasticity. In particular, a class of models that has received considerable attention is the ARFIMA–GARCH (generalized autoregressive conditional heteroskedastic) process, see for example Ling and Li (1997). In this model, the returns have long memory and the noise has a conditional heteroskedasticity structure. A related class of interesting models is the extension of the ARCH(p) processes first introduced by Engle (1982) to the ARCH(∞) models to encompass the longer dependence observed in many squared financial series. On the other hand, extensions of the stochastic volatility processes to the long-memory case

have produced the so-called long-memory stochastic volatility models (LMSV). In this section, we discuss briefly some of these well-known econometric models.

8.1. ARFIMA–GARCH Model

An ARFIMA(p,d,q)–GARCH(r,s) process is defined by the discrete-time equation

$$\begin{aligned}\Phi(B)(1-B)^d(y_t-\mu) &= \Theta(B)\varepsilon_t \\ \varepsilon|\mathscr{F}_{t-1} &\sim N(0,\ h_t) \\ h_t &= \alpha_0+\sum_{i=1}^{r}\alpha_i\varepsilon_t^2+\sum_{j=1}^{s}\beta_j h_{t-j}\end{aligned}$$

where $\mathscr{F}_{t-1}$ is the σ-algebra generated by the past observations $y_{t-1}, y_{t-2}, \ldots$.

Most econometric models dealing with long memory and heteroskedastic behaviors are non-linear, in the sense that the noise sequence is not necessarily independent. An approximate MLE $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the conditional log-likelihood

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2n}\sum_{t=1}^{n}\left\{\log\ h_t+\frac{\varepsilon_t^2}{h_t}\right\} \tag{33}$$

The asymptotic behavior of this estimate was formally established by Ling and Li (1997). Let $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2)'$, where $\boldsymbol{\theta}_1=(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, d)'$ is the parameter vector involving the ARFIMA components and $\boldsymbol{\theta}_2=(\alpha_0, \ldots, \alpha_r, \beta_1, \ldots, \beta_s)'$ is the parameter vector containing the GARCH component. The following result correspond to Theorem 3.2 of Ling and Li (1997).

Theorem 9. Let $\hat{\boldsymbol{\theta}}_n$ be the value that maximizes the conditional log-likelihood function (33). Then, under some regularity conditions, $\hat{\boldsymbol{\theta}}_n$ is a consistent estimate and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\to N(0,\ \Omega^{-1})$, as $n\to\infty$, where $\Omega=\mathrm{diag}(\Omega_1,\Omega_2)$ with

$$\Omega_1 = E\left[\frac{1}{h_t}\frac{\partial\varepsilon_t}{\partial\boldsymbol{\theta}_1}\frac{\partial\varepsilon_t}{\partial\boldsymbol{\theta}_1'}+\frac{1}{2h_t^2}\frac{\partial h_t}{\partial\boldsymbol{\theta}_1}\frac{\partial h_t}{\partial\boldsymbol{\theta}_1'}\right]$$

and

$$\Omega_2 = E\left[\frac{1}{2h_t^2}\frac{\partial h_t}{\partial \boldsymbol{\theta}_2}\frac{\partial h_t}{\partial \boldsymbol{\theta}_2'}\right]$$

8.2. Arch-Type Models

The ARFIMA–GARCH process described in the previous subsection is adequate for modeling long-range dependence in returns of financial time series. However, as described by Rosenblatt (1961) and Palma and Zevallos (2004), the squares of an ARFIMA–GARCH process have only *intermediate memory* for $d \in (0, 1/4)$. In fact, for any $d \in (0, 1/2)$, the ACF of the squared series behaves like $k^{2\tilde{d}-1}$, where k denotes the k-*th* lag and $\tilde{d} = 2d - 1/2$. Consequently, the long-memory parameter of the squared series $\tilde{d}$ is always smaller than the long-memory parameter of the original series d, i.e. $\tilde{d} < d$ for $d < 1/2$.

Since in many financial applications the squared returns have the same or greater level of autocorrelation, the theoretical reduction in the memory that affects the squares of an ARFIMA–GARCH process may not be appropriate in practice. This situation leads one to consider other classes of processes to model the dependence of the squared returns directly. For instance, Robinson (1991) proposed the following extension of the ARCH(p) introduced by Engle (1982),

$$y_t = \sigma_t \xi_t$$
$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j y_{t-j}^2$$

which can be formally written as

$$y_t^2 = \alpha_0 + v_t + \sum_{j=1}^{\infty} \alpha_j y_{t-j}^2 \tag{34}$$

where $\sigma_t^2 = E[y_t^2 | y_{t-1}, y_{t-2}, \ldots]$, $\quad v_t = y_t^2 - \sigma_t^2$ is a martingale difference sequence, $\{\xi_t\}$ a sequence of independent and identically distributed random variables and α_0 a positive constant, cf. Eqs. (1.31), (1.33) and (1.35) of Robinson (2003), respectively.

When the coefficients $\{\alpha_j\}$ of (34) are specified by an ARFIMA(p,d,q) model, the resulting process corresponds to the (fractionally integrated

GARCH) FIGARCH(p,d,q) model which is defined by

$$\Phi(B)(1-B)^d y_t^2 = \omega + \Theta(B)v_t$$

where ω is a positive constant, cf. Baillie, Bollerslev, & Mikkelsen, (1996). As noted by Karanasos, Psaradakis, & Sola (2004), this process is strictly stationary and ergodic but not square integrable.

8.2.1. Estimation

Consider the quasi log-likelihood function

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left\{\log\ \sigma_t^2 + \frac{\varepsilon_t^2}{\sigma_t^2}\right\} \tag{35}$$

where $\boldsymbol{\theta} = (\omega, d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_1)$. A QMLE $\hat{\boldsymbol{\theta}}_n$ can be obtained by maximizing (35). But, even though this estimation approach has been widely used in many practical applications, to the best of our knowledge, asymptotic results for these estimators remain an open issue. For a recent study about this problem, see for example Caporin (2002).

8.3. Stochastic Volatility

Stochastic volatility models have been addressed by Harvey, Ruiz, and Shephard (1994), Ghysels et al. (1996) and Breidt et al. (1998), among others. These processes are defined by

$$r_t = \sigma_t \xi_t$$

and

$$\sigma_{\mathrm{t}} = \sigma\ \exp(v_t/2) \tag{36}$$

where $\{\xi_t\}$ is a independent, identically distributed sequence with mean zero and variance one and $\{v_t\}$ is a stationary process independent of $\{\xi_t\}$. In particular, $\{v_t\}$ can be specified as a long-memory ARFIMA(p,d,q) process. The resulting process is called LMSV model.

From (36), we can write

$$\log(r_t^2) = \log(\sigma_t^2) + \log(\xi_t^2)$$
$$\log(\sigma_t^2) = \log(\sigma^2) + v_t$$

Let $y_t = \log(r_t^2)$, $\mu = \log(\sigma^2) + E[\log(\xi_t^2)]$ and $\varepsilon_t = \log(\xi_t^2) - E[\log(\xi_t^2)]$. Then

$$y_t = \mu + v_t + \varepsilon_t \tag{37}$$

Consequently, the transformed process $\{y_t\}$ corresponds to a stationary long-memory process plus an independent noise.

The ACF of (37) is given by

$$\gamma_y(\kappa) = \gamma_v(\kappa) + \sigma_\varepsilon^2 \delta_0(\kappa)$$

where $\delta_0(k) = 1$ for $k = 0$ and $\delta_0(k) = 0$ otherwise. Furthermore, the spectral density of $\{y_t\}$, f_y, is given by

$$f_y(\omega) = f_v(\omega) + \frac{\sigma_\varepsilon^2}{2\pi}$$

where f_v is the spectral density of the long-memory process $\{v_t\}$.

In particular, if the process $\{v_t\}$ is an ARFIMA(p,d,q) model

$$\Phi(B)(1 - B)^d v_t = \Theta(B)\eta_t \tag{38}$$

and $\boldsymbol{\theta} = (d, \sigma_\eta^2, \sigma_\varepsilon^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ is the parameter vector that specifies model (38), then the spectral density is given by

$$f_{\boldsymbol{\theta}}(\omega) = \frac{\sigma_\eta^2}{2\pi} \frac{\left|\Theta(\exp(i\omega)\right|^2}{\left|1 - \exp(i\omega)\right|^{2d} \left|\Phi(\exp(i\omega)\right|^2} + \frac{\sigma_\varepsilon^2}{2\pi}$$

Breidt et al. (1998) consider the estimation of the parameter $\boldsymbol{\theta}$ by means of the spectral-likelihood estimator obtained by minimizing

$$\mathscr{L}_7(\boldsymbol{\theta}) = \frac{2\pi}{n} \sum_{j=1}^{n/2} \left\{ \log f_{\boldsymbol{\theta}}(\omega_j) + \frac{I(\omega_j)}{f_{\boldsymbol{\theta}}(\omega_j)} \right\}$$

where $f_{\boldsymbol{\theta}}(\omega)$ is given by (39).

Let $\hat{\boldsymbol{\theta}}$ be the value that minimizes $\mathscr{L}_7(\boldsymbol{\theta})$ over the parameter space Θ. Breidt et al. (1998) prove the following result.

Theorem 10. Assume that the parameter vector $\boldsymbol{\theta}$ is an element of the compact parameter space Θ and assume that $f_{\boldsymbol{\theta}_1} = f_{\boldsymbol{\theta}_2}$ implies that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. Let $\boldsymbol{\theta}_0$ be the true parameter value. Then $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability as $n \to \infty$.

Other estimation procedures for LMSV using state space systems can be found in Chan and Petris (2000) and Section 11 of Chan (2002).

Table 5. Quasi ML Estimation of Long-Memory Stochastic Volatility Models with an ARFIMA(0, d, 0) Specification for v_t, for Different Values of d.

d	$\hat{d}$	$\hat{\sigma}_\eta$	S.D.($\hat{d}$)	S.D.($\hat{\sigma}_\eta$)
0.10	0.0868	9.9344	0.0405	0.4021
0.25	0.2539	10.0593	0.0400	0.4199
0.40	0.4139	10.1198	0.0415	0.3773

8.4. Numerical Experiments

The finite sample performance of the spectral-likelihood estimator is analyzed here by means of Monte Carlo simulations. The model investigated is the LMSV with an ARFIMA(0,d,0) structure, $\sigma_\varepsilon = \pi/\sqrt{2}, \xi_t$ follows a standard normal distribution, $\sigma_\eta = 10$ and the sample size is $n = 400$. The results displayed in Table 5 are based on 1,000 replications.

From Table 5, observe that estimates of both the long-memory parameter d and the scale parameter σ_η are close to their true values. On the other hand, the standard deviations of $\hat{d}$ and $\hat{\sigma}_\eta$ seem to be similar for all the values of d simulated. However, to the best of our knowledge there are no formally established results for the asymptotic distribution of these QMLE yet.

9. SUMMARY

In this article, a number of estimation techniques for long-memory time series have been reviewed together with their corresponding asymptotic results. Finite sample behaviors of these techniques were studied through Monte Carlo simulations. It is found that they are relatively comparable in terms of finite sample performance. However, in situations like missing data or long-memory seasonal time series, some approaches such as the MLE or truncated MLE seems to be more efficient than their spectral domain counterparts such as the Whittle approach.

Clearly, long-memory time series is an exciting and important topic in econometrics as well as many other disciplines. This article does not attempt to cover all of the important aspects of this exciting field. Interested readers may find many actively pursued topics in this area in the recent monograph of Robinson (2003). It is hoped that this article offers a focused and practical introduction to the estimation of long-memory time series.

ACKNOWLEDGMENT

We would like to thank an anonymous referee for helpful comments and Mr. Ricardo Olea for carrying out some of the simulations reported in this paper. This research is suppored in part by HKSAR-RGC Grants CUHK 4043/02P, 400305 and 2060268 and Fondecyt Grant 1040934. Part of this research was conducted when the second author was visiting the Institute of Mathematical Sciences at the Chinese University of Hong Kong. Support from the IMS is gratefully acknowledged.

REFERENCES

Ammar, G. S. (1996). Classical foundations of algorithms for solving positive definite Toeplitz equations. *Calcolo*, *33*, 99–113.

Arteche, J., & Robinson, P. M. (2000). Semiparametric inference in seasonal and cyclical long memory processes. *Journal of Time Series Analysis*, *21*, 1–25.

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

Beran, J. (1994a). On a class of *M*-estimators for Gaussian long-memory models. *Biometrika*, *81*, 755–766.

Beran, J. (1994b). *Statistics for long-memory processes*. New York: Chapman and Hall.

Bertelli, S., & Caporin, M. (2002). A note on calculating autocovariances of long-memory processes. *Journal of Time Series Analysis*, *23*, 503–508.

Bhansali, R. J., & Kokoszka, P. S. (2003). Prediction of long-memory time series. In: P. Doukhan, G. Oppenheim & M. Taqqu (Eds), *Theory and applications of long-range dependence* (pp. 355–367). Boston: Birkhäuser.

Bondon, P., & Palma, W. (2005). Prediction of strongly dependent time series. Preprint, Pontificia Universidad Catolica de Chile, Santiago, Chile.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis, forecasting and control*. Englewood Cliffs: Prentice-Hall.

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

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods*. New York: Springer-Verlag.

Caporin, M. (2002). Long memory conditional heteroskedasticity and second order causality. Ph.D. Dissertation, Universita de Venezia, Venezia.

Chan, N. H. (2002). *Time series, applications to finance*. Wiley Series in Probability and Statistics. New York: Wiley

Chan, N. H., & Palma, W. (1998). State space modeling of long-memory processes. *The Annals of Statistics*, *26*, 719–740.

Chan, N. H. & Petris, G. (2000). Recent developments in heteroskedastic financial series. In: W. S. Chan, W. K. Li & H. Tong (Eds), *Statistics and finance* (pp. 169–184). (Hong Kong, 1999). London: Imperial College Press.

Chung, C. F. (1996). A generalized fractionally integrated autoregressive moving-average process. *Journal of Time Series Analysis*, *17*, 111–140.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *The Annals of Statistics*, *17*, 1749–1766.

Durbin, J. (1960). The fitting of time series models. *Review of The International Statistical Institute*, *28*, 233–244.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*, 987–1007.

Fox, R., & Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Annals of Statistics*, *14*, 517–532.

Ghysels, E., Harvey, A. C., & Renault, E. (1996). Stochastic volatility. In: G. S. Maddala & C. R. Rao (Eds), *Statistical methods in finance* (Vol. 14 of Handbook of Statistics. pp. 119–191). Amsterdam: North-Holland.

Giraitis, L., Hidalgo, J., & Robinson, P. M. (2001). Gaussian estimation of parametric spectral density with unknown pole. *The Annals of Statistics*, *29*, 987–1023.

Giraitis, L., & Leipus, R. (1995). A generalized fractionally differencing approach in long-memory modeling. *Lietuvos Matematikos Rinkinys*, *35*, 65–81.

Giraitis, L., & Surgailis, D. (1990). A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate. *Probability Theory and Related Fields*, *86*, 87–104.

Gradshteyn, I. S., & Ryzhik, I. M. (2000). *Table of integrals, series, and products*. San Diego, CA: Academic Press Inc.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, *1*, 15–29.

Gray, H. L., Zhang, N. F., & Woodward, W. A. (1989). On generalized fractional processes. *Journal of Time Series Analysis*, *10*, 233–257.

Grenander, U., & Szegö, G. (1958). Toeplitz forms and their applications. California monographs in mathematical sciences. Berkeley: University of California Press.

Harvey, A. C., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, *61*, 247–265.

Hassler, U. (1994). (Mis)specification of long memory in seasonal time series. *Journal of Time Series Analysis*, *15*, 19–30.

Hassler, U., & Wolters, J. (1995). Long memory in inflation rates: International evidence. *Journal of Business and Economic Statistics*, *13*, 37–45.

Haslett, J., & Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *Journal of Applied Statistics*, *38*, 1–50.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, *68*, 165–176.

Karanasos, M., Psaradakis, Z., & Sola, M. (2004). On the autocorrelation properties of long-memory GARCH processes. *Journal of Time Series Analysis*, *25*, 265–281.

Levinson, N. (1947). The Wiener RMS (root mean square) error criterion in filter design and prediction. *Journal of Mathematical Physics of the Massachusetts Institute of Technology*, *25*, 261–278.

Li, W. K., & McLeod, A. I. (1986). Fractional time series modelling. *Biometrika*, *73*, 217–221.

Ling, S., & Li, W. K. (1997). On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *Journal of the American Statistical Association*, *92*, 1184–1194.

Montanari, A., Rosso, R., & Taqqu, M. S. (2000). A seasonal fractional ARIMA model applied to Nile River monthly flows at Aswan. *Water Resources Research*, *36*, 1249–1259.

Ooms, M. (1995). *Flexible seasonal long memory and economic time series*. Rotterdam. Technical report Econometric Institute, Erasmus University.

Palma, W., & Chan, N. H. (1997). Estimation and forecasting of long-memory processes with missing values. *Journal of Forecasting*, *16*, 395–410.

Palma, W., & Chan, N. H. (2005). Efficient estimations of seasonal long range dependent processes. *Journal of Time Series Analysis*, *26*, 863–892.

Palma, W., & Del Pino, G. (1999). Statistical analysis of incomplete long-range dependent data. *Biometrika*, *86*, 965–972.

Palma, W., & Zevallos, M. (2004). Analysis of the correlation structure of square time series. *Journal of Time Series Analysis*, *25*, 529–550.

Porter-Hudak, S. (1990). An application of the seasonal fractionally differenced model to the monetary aggregates. *Journal of the American Statistical Association, Applications and Case Studies*, *85*, 338–344.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN*. Cambridge: Cambridge University Press.

Ray, B. K. (1993). Long-range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA model. *International Journal of Forecasting*, *9*, 255–269.

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. (1995). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, *23*, 1630–1661.

Robinson, P. M. (2003). *Time series with long memory*. Oxford: Oxford University Press.

Rosenblatt, M. (1961). Independence and dependence. In: J. Neyman (Ed.), *Proceeding of the 4th Berkeley Symposium On Mathematics Statistics and Probability* (Vol. II, pp. 431–443). Berkeley, CA: University of California Press.

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications*. New York: Springer-Verlag.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics*, *53*, 165–188.

Velasco, C., & Robinson, P. M. (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *Journal of the American Statistical Association*, *95*, 1229–1243.

Whittle, P. (1951). *Hypothesis testing in time series analysis*. New York: Hafner.

Woodward, W. A., Cheng, Q. C., & Gray, H. L. (1998). A *k*-factor GARMA long-memory model. *Journal of Time Series Analysis*, *19*, 485–504.

Yajima, Y. (1985). On estimation of long-memory time series models. *The Australian Journal of Statistics*, *27*, 303–320.

This page intentionally left blank

BOOSTING-BASED FRAMEWORKS IN FINANCIAL MODELING: APPLICATION TO SYMBOLIC VOLATILITY FORECASTING

Valeriy V. Gavrishchaka

ABSTRACT

Increasing availability of the financial data has opened new opportunities for quantitative modeling. It has also exposed limitations of the existing frameworks, such as low accuracy of the simplified analytical models and insufficient interpretability and stability of the adaptive data-driven algorithms. I make the case that boosting (a novel, ensemble learning technique) can serve as a simple and robust framework for combining the best features of the analytical and data-driven models. Boosting-based frameworks for typical financial and econometric applications are outlined. The implementation of a standard boosting procedure is illustrated in the context of the problem of symbolic volatility forecasting for IBM stock time series. It is shown that the boosted collection of the generalized autoregressive conditional heteroskedastic (GARCH)-type models is systematically more accurate than both the best single model in the collection and the widely used GARCH(1,1) model.

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 123–151
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20024-5

1. INTRODUCTION

Increasing availability of the high-frequency and multisource financial data has opened new opportunities for quantitative modeling and problem solving in a wide range of financial and econometric applications. These include such challenging problems as forecasting of the financial time series and their volatilities, portfolio strategy discovery and optimization, valuation of complex derivate instruments, prediction of rare events (e.g., bankruptcy and market crashes), and many others.

Such data availability also allows more accurate testing and validation of the existing models and frameworks to expose their limitations and/or confirm their advantages. For example, analytical and simple parametric models are usually clear and stable, but lack sufficient accuracy due to simplified assumptions. Adaptive data-driven models (mostly non-parametric or semi-parametric) can be very accurate in some regimes, but lack consistent stability and explanatory power (''black-box'' feature). Moreover, the multivariate nature of many problems and data non-stationarity results in ''dimensionality curse'' (Bishop, 1995) and incompleteness of the available training data that is prohibitive for a majority of the machine learning and statistical algorithms.

Recently, the large margin classification techniques emerged as a practical result of the statistical learning theory, in particular, support vector machines (SVMs) (Vapnik, 1995, 1998; Cristianini & Shawe-Taylor, 2000). A large margin usually implies good generalization performance. The margin is the distance of the example to the class separation boundary. In contrast to many existing machine learning and statistical techniques, a large margin classifier generates decision boundaries with large margins to almost all training examples. This results in superior generalization ability on out-of-sample data.

SVMs for classification and regression have been successfully applied to many challenging practical problems. Recent successful applications of the SVM-based adaptive systems include market volatility forecasting (Gavrishchaka & Ganguli, 2003), financial time series modeling (Van Gestel et al., 2001), bankruptcy prediction (Fan & M. Palaniswami, 2000), space weather forecasting (Gavrishchaka & Ganguli, 2001a, b), protein function classification (Cai, Wang, Sun, & Chen, 2003), cancer diagnostics (Chang, Wu, Moon, Chou, & Chen, 2003), face detection and recognition (Osuna, Freund, & Girosi, 1997) as well as many other financial, scientific, engineering, and medical applications. In most cases, SVMs demonstrated superior results compared to other advanced machine learning and statistical techniques, including different types of neural networks (NNs).

Although SVMs are significantly more tolerant to data dimensionality and incompleteness (Cristianini & Shawe-Taylor, 2000; Gavrishchaka & Ganguli, 2001a, b), they can still fail in many important cases where available data does not allow generating stable and robust data-driven models. Moreover, SVM framework, like many other machine learning algorithms (including NNs), largely remains to be a "black box" with its limited explanatory power. Also, it is not easy to introduce a priori problem-specific knowledge into the SVM framework, except for some flexibility in the problem-dependent choice of the kernel type.

One of the machine learning approaches to compensate for the deficiency of the individual models is to combine several models to form a committee (e.g., Bishop, 1995; Hastie, Tibshirani, & Friedman, 2001; Witten & Frank, 2000). The committee can compensate limitations of the individual models due to both incomplete data and specifics of the algorithms (e.g., multiple local minima in the NN error surface). A number of different ensemble learning (model combination) techniques to build optimal committees have been proposed over the years in several research communities (e.g., Granger, 1989; Clemen, 1989; Drucker et al., 1994; Hastie, Tibshirani, & Friedman, 2001, and references therein). The work of Bates and Granger (1969) is considered to be the first seminal work on forecast combining (model mixing) in econometrics and financial forecasting. In the early machine learning literature, the same area of research was called "evidence combination" (e.g., Barnett, 1981).

Recently, there was a resurgence of interest in model combination in machine learning community (e.g., Schapire, 1992; Drucker et al., 1994; Dietterich, 2000; Ratsch, 2001). Modern research is mainly focused on the novel techniques that are suited for challenging problems with such features as a large amount of noise, limited number of training data, and high-dimensional patterns. These types of problems often arise in financial applications dealing with high-frequency and heterogeneous data (e.g., Gavrishchaka & Ganguli, 2003 and references therein), bioinformatics and computational drug discovery (e.g., Scholkopf, Tsuda, & Vert, 2004), space weather forecasting (e.g., Gleisner & Lundstedt, 2001; Gavrishchaka & Ganguli, 2001a, b), medical diagnostic systems (e.g., Gardner, 2004), and many others. Many modern ensemble learning algorithms perform a special manipulation with the training data set to compensate for the data incompleteness and its numerous consequences. Among them are bagging, cross-validating committees, and boosting (e.g., Drucker et al., 1994; Ratsch, 2001; Hastie et al., 2001).

The advantage of the ensemble learning approach is not only the possibility of the accuracy and stability improvement, but also its ability to

combine a variety of models: analytical, simulation, and data-driven. This latter feature can significantly improve explanatory power of the combined model if building blocks are sufficiently simple and well-understood models. However, ensemble learning algorithms can be susceptible to the same problems and limitations as standard machine learning and statistical techniques. Therefore, the optimal choice of both the base model pool and ensemble learning algorithms with good generalization qualities and tolerance to data incompleteness and dimensionality is very important.

A very promising ensemble learning algorithm that combines many desirable features is boosting (Valiant, 1984; Schapire, 1992; Ratsch, 2001). Boosting and its specific implementations such as AdaBoost (Freund & Schapire, 1997) have been actively studied and successfully applied to many challenging classification problems (Drucker, Schapire, & Simard, 1993; Drucker et al., 1994; Schwenk & Bengio, 1997; Opitz & Maclin, 1999). Recently, boosting has been successfully applied to such practical problems as document routing (Iyer et al., 2000), text categorization (Schapire & Singer, 2000), face detection (Xiao, Zhu, & Zang, 2003), pitch accent prediction (Sun, 2002), and others.

One of the main features that sets boosting aside from other ensemble learning frameworks is that it is a large margin classifier similar to SVM. Recently, connection between SVM and boosting was rigorously proven (Schapire et al., 1998). This ensures superior generalization ability and better tolerance to incomplete data compared to other ensemble learning techniques. Similar to SVM, the boosting generalization error does not directly depend on the dimensionality of the input space (Ratsch, 2001). Therefore, boosting is also capable of working with high-dimensional patterns. The other distinctive and practically important feature of boosting is its ability to produce a powerful classification system starting with just a ''weak'' classifier as a base model (Ratsch, 2001).

In most cases discussed in the literature, boosting is used to combine data-driven models based on machine learning algorithms such as classification trees, NNs, etc. In this article, I will stress out advantages of the boosting framework that allows an intelligent combination of the existing analytical and other simplified and parsimonious models specific to the field of interest. In this way, one can combine clarity and stability typical for analytical and parsimonious parametric models with a good accuracy usually achieved only by the best adaptive data-driven algorithms. Moreover, since the underlying components are well-understood and accepted models, the obtained ensemble is not a pure ''black box.''

In the next two sections, I provide a short overview of the boosting key features and its relation to existing approaches as well as the description of

the algorithm itself. After that, potential applications of boosting in quantitative finance and financial econometrics are outlined. Finally, a realistic example of the successful boosting application to symbolic volatility forecasting is given. In particular, boosting is used to combine different types of GARCH models for the next day threshold forecasting of the IBM stock volatility. Ensemble obtained with a regularized AdaBoost is shown to be consistently superior to both single best model and GARCH(1,1) model on both training (in-sample) and test (out-of-sample) data.

2. BOOSTING: THE MAIN FEATURES AND RELATION TO OTHER TECHNIQUES

Since the first formulation of the adaptive boosting as a novel ensemble learning (model combination) algorithm (Schapire, 1992), researchers from different fields demonstrated its relation to different existing techniques and frameworks. Such an active research was inspired by the boosting robust performance in a wide range of applications (Drucker et al., 1993, 1994; Schwenk & Bengio, 1997; Opitz & Maclin, 1999; Schapire & Singer, 2000; Xiao et al., 2003). The most common conclusion in machine learning community is that boosting represents one of the most robust ensemble learning methods. Computational learning theorists also proved boosting association with the theory of margins and SVMs that belong to an important class of large-margin classifiers (Schapire et al., 1998).

Recent research in statistical community is showing that boosting can also be viewed as an optimization algorithm in function space (Breiman, 1999). Statisticians consider boosting as a new class of learning algorithms that Friedman named ''gradient machines'' (Friedman, 1999), since boosting performs a stage wise greedy gradient descent. This relates boosting to particular additive models and matching pursuit known within the statistics literature (Hastie et al., 2001). There are also arguments that the original class of model mixing procedures are not in competition with boosting but rather can coexist inside or outside a boosting algorithm (Ridgeway, 1999).

In this section, I review boosting key features in the context of its relation to other techniques. Since the structure and the most practical interpretation of the boosting algorithm naturally relates it to ensemble learning techniques, comparison of boosting with other approaches for model combination will bo my main focus. Conceptual similarity and differences with other algorithms will be demonstrated from several different angles whenever possible.

The practical value of model combination is exploited by practitioners and researchers in many different fields. In one of his papers, Granger (1989) summarizes the usefulness of combining forecasts: ''The combination of forecasts is a simple, pragmatic, and sensible way to possibly produce better forecasts.'' The basic idea of ensemble learning algorithms including boosting is to combine relatively simple base hypotheses (models) for the final prediction. The important question is why and when an ensemble is better than a single model.

In machine learning literature, three broad reasons for possibility of good ensembles' construction are often mentioned (e.g., Dietterich, 2000). First, there is a pure statistical reason. The amount of training data is usually too small (data incompleteness) and learning algorithms can find many different models (from model space) with comparable accuracy on the training set. However, these models capture only certain regimes of the whole dynamics or mapping that becomes evident in out-of-sample performance.There is also a computational reason related to the learning algorithm specifics such as multiple local minima on the error surface (e.g., NNs and other adaptive techniques). Finally, there is a representational reason when the true model cannot be effectively represented by a single model from a given set even for the adequate amount of training data. Ensemble methods have a promise of reducing these key shortcomings of standard learning algorithms and statistical models.

One of the quantitative and explanatory measures for the analysis and categorization of the ensemble learning algorithms is an error diversity (e.g., Brown, Wyatt, Harris, & Yao, 2005 and references therein). In particular, the ambiguity decomposition (Krogh & Vedelsby, 1995) and bias variance–covariance decomposition (Geman, Bienenstack, & Dourstat, 1992) provide a quantification of diversity for linearly weighted ensembles by connecting it back to an objective error criterion: mean-squared error. Although these frameworks have been formulated for regression problems, they can also be useful for the conceptual analysis of the classifier ensembles (Brown et al., 2005).

For simplicity, I consider only ambiguity decomposition that is formulated for convex combinations and is a property of an ensemble trained on a single dataset. Krogh and Vedelsby (1995) proved that at a single data point the quadratic error of the ensemble is guaranteed to be less than or equal to the average quadratic error of the base models:

$$(f-y)^2=\sum_t w_t(f_t-y)^2-\sum_t w_t(f_t-f)^2 \tag{1}$$

where f is a convex combination ($\sum_t w_t = 1$) of the base models f_t:

$$f = \sum_t w_t f_t \tag{2}$$

This result directly shows the effect due to error variability of the base models. The first term in (1) is the weighted average error of the base models. The second is the ambiguity term, measuring the amount of variability among the ensemble member answers for a considered pattern. Since this term is always positive, it is subtractive from the first term. This means that the ensemble is guaranteed lower error than the average individual error.

The larger the ambiguity term (i.e., error diversity), the larger is the ensemble error reduction. However, as the variability of the individual models rises, so does the value of the first term. Therefore, in order to achieve the lowest ensemble error, one should get the right balance between error diversity and individual model accuracy (similar to the bias-variance compromise for a single model). The same is conceptually true for the classification ensembles (Hansen & Salamon, 1990).

The above discussion offers some clues about why and when the model combination can work in practice. However, the most important practical question is how to construct robust and accurate ensemble learning methods. In econometric applications, the main focus is usually on equal weight and Bayesian model combination methods (e.g., Granger, 1989; Clemen, 1989). These methods provide a simple and well-grounded procedure for the combination of the base models chosen by a practitioner or a researcher. The accuracy of the final ensemble crucially depends on the accuracy and diversity of the base models. In some cases, a priori knowledge and problem-dependent heuristics can help to choose an appropriate pool of the base models.

However, equal weight and Bayesian combination methods do not provide any algorithmic procedures for the search, discovery, and building of the base models that are suitable for the productive combination in an ensemble. Such procedures would be especially important in complex high-dimensional problems with limited a priori knowledge and lack of simple heuristics. Without such algorithms for the automatic generation of the models with significant error diversity, it could be difficult or impossible to create powerful and compact ensembles from the simple base models.

In modern machine learning literature, the main focus is on the ensemble learning algorithms suited for challenging problems dealing with a large amount of noise, limited number of training data, and high-dimensional patterns (e.g., Ratsch, 2001; Buhlmann, 2003). Several modern ensemble

learning techniques relevant for these types of applications are based on training data manipulation as a source of base models with significant error diversity. These include such algorithms as bagging (''bootstrap aggregation''), cross-validating committees, and boosting (e.g., Ratsch, 2001; Witten & Frank, 2000; Hastie et al., 2001, and references therein).

Bagging is a typical representative of ''random sample'' techniques in ensemble construction. In bagging, instances are randomly sampled, with replacement, from the original training dataset to create a bootstrap set with the same size (e.g., Witten & Frank, 2000; Hastie et al., 2001). By repeating this procedure, multiple training data sets are obtained. The same learning algorithm is applied to each data set and multiple models are generated. Finally, these models are linearly combined as in (2) with equal weights. Such combination reduces variance part of the model error and instability caused by the training set incompleteness.

Unlike basic equal weight and Bayesian combination methods, bagging offers a direct procedure to build base models with potentially significant error diversity (the last term in (1)). Recently, a number of successful econometric applications of bagging have been reported and compared with other model combination techniques (e.g., Kilian & Inoue, 2004).

Bagging exploits the instability inherent in learning algorithms. For example, it can be successfully applied to the NN-based models. However, bagging is not efficient for the algorithms that are stable, i.e., whose output is not sensitive to small changes in the input (e.g., parsimonious parametric models). Bagging is also not suitable for a consistent bias reduction.

Intuitively, combining multiple models helps when these models are significantly different from one another and each one treats a reasonable portion of the data correctly. Ideally, the models should complement one another, each being an expert in a part of the domain where performance of other models is not satisfactory. The boosting method for combining multiple models exploits this insight by explicitly seeking and/or building models that complement one another (Valiant, 1984; Schapire, 1992; Ratsch, 2001).

Unlike bagging, boosting is iterative. Whereas in bagging individual models are built separately, in boosting, each new model is influenced by the performance of those built previously. Boosting encourages new models to become experts for instances handled incorrectly by earlier ones. Final difference is that boosting weights obtained models by their performance, i.e., weights are not equal as in bagging. Unlike bagging and similar ''random sample'' techniques, boosting can reduce both bias and variance parts of the model error.

The initial motivation for boosting was a procedure that combines the outputs of many ''weak'' classifiers to produce a powerful committee (Valiant, 1984; Schapire, 1992; Ratsch, 2001). The purpose of boosting is to sequentially apply the weak classification algorithm to repeatedly modified versions of data, thereby producing sequence of weak classifiers. The predictions from all of them are then combined through a weighted majority vote to produce the final prediction. As iterations proceed, observations that are difficult to correctly classify receive an ever-increasing influence through larger weight assignment. Each successive classifier trained on the weighted error function is thereby forced to concentrate on those training observations that are missed by previous ones in the sequence.

Empirical comparative studies of different ensemble methods often indicate superiority of boosting over other techniques (Drucker et al., 1993, 1994; Schwenk & Bengio, 1997; Opitz & Maclin, 1999). Using probably approximately correct (PAC) theory, it was shown that if the base learner is just slightly better than random guessing, AdaBoost is able to construct ensemble with arbitrary high accuracy (Valiant, 1984). Thus, boosting can be effectively used to construct powerful ensembles from the very simplistic ''rules of thumb'' known in the considered field.

The distinctive features of boosting can also be illustrated through the error diversity point of view. In constructing the ensemble, the algorithm could either take information about error diversity into account or not. In the first case, the algorithm explicitly tries to optimize some measure of diversity during building the ensemble. This allows to categorize ensemble learning techniques as explicit and implicit diversity methods. While implicit methods rely on randomness to generate diverse trajectories in the model space, explicit methods deterministic ally choose different paths in the space.

In this context, bagging is an implicit method. It randomly samples the training patterns to produce different sets for each model and no measurement is taken to ensure emergence of the error diversity. On the other hand, boosting is an explicit method. It directly manipulates the training data distributions through specific weight changes to ensure some form of diversity in the set of base models. As mentioned before, equal weight and Bayesian model combination methods do not provide any direct algorithmic tools to manipulate error diversity of the base models.

Boosting has also a tremendous advantage over other ensemble methods in terms of interpretation. At each iteration, boosting trains new models or searches the pool of existing models that are complementary to the already chosen models. This often leads to the very compact but accurate ensemble with a clear interpretation in the problem-specific terms. This could also

provide an important byproduct in the form of automatic discovery of the complementary models that may have value of their own in the area of application.

Boosting also offers a flexible framework for the incorporation of other ensemble learning techniques. For example, at each boosting iteration, instead of taking just one best model, one can use mini-ensemble of models that are chosen or built according to some other ensemble learning technique. From my empirical experience with boosting. I find it useful to form an equal weight ensemble of several comparable best models at each boosting iteration. Often, this leads to the superior out-of-sample performance compared to the standard boosted ensemble. Of course, there are many different ways to combine boosting with other ensemble techniques that could be chosen according to the specifics of the application.

So far, the main boosting features have been illustrated through its relation to other ensemble learning methods. However, one of the most distinctive and important features of boosting relates it to the large margin classifiers. It was found that boosted ensemble is a large margin classifier and that SVM and AdaBoost are intimately related (Schapire et al., 1998). Both boosting and SVM can be viewed as attempting to maximize the margin, except that the norm used by each technique and optimization procedures are different (Ratsch, 2001).

It is beyond the scope of this paper to review the theory of margins and structural risk minimization that is the foundation of the large margin classifiers including SVM (Vapnik, 1995, 1998). However, it should be mentioned that the margin of the data example is its minimal distance (in the chosen norm) to the classifier decision boundary. Large margin classifiers attempt to find decision boundary with large (or maximum) margin for all training data examples. Intuitively, this feature should improve the generalization ability (i.e., out-of-sample performance) of the classifier. Rigorous treatment of this topic can be found in (Vapnik, 1995, 1998; Ratsch, 2001).

Relation to the large margin classifiers partially explains robust generalization ability of boosting. It also clarifies its ability to work with high-dimensional patterns. Many traditional statistical and machine learning techniques often have the problem of "dimensionality curse" (Bishop, 1995), i.e., inability of handling high-dimensional inputs. The upper bound of the boosting generalization error depends on the margin and not on the dimensionality of the input space (Ratsch, 2001). Hence, it is easy to handle high-dimensional data and learn efficiently if the data can be separated with large margin. It has been shown that boosting superiority over other techniques is more pronounced in high-dimensional problems (e.g., Buhlmann, 2003).

3. ADAPTIVE BOOSTING FOR CLASSIFICATION

In this section, the technical details of a typical boosting algorithm and its properties are presented. For clarity, I will describe boosting only for two-class classification problem. A standard formalization of this task is to estimate a function $f : X \to Y$, where X is the input domain (usually R^n) and $Y = \{-1, +1\}$ is the set of possible class labels. An example (x_i, y_i) is assigned to the class $+1$ if $f(x) \geq 0$ and to the class -1 otherwise. Optimal function f from a given function set F is found by minimizing an error function (empirical risk) calculated on a training set $S = \{(x_1, y_1), ..., (x_N, y_N)\}$ with a chosen loss function $g(y, f(x))$:

$$\varepsilon[f, S] = \frac{1}{N}\sum_{n=1}^{N} g(y_n, f(x_n)) \tag{3}$$

Unlike the typical choice of squared loss in regression problems, here one usually considers the 0/1 loss:

$$g(y, f(x)) = I(-yf(x)) \tag{4}$$

where $I(z) = 0$ for $z < 0$ and $I(z) = 1$ otherwise. In a more general case, a cost measure of interest can be introduced through the multiplication of $g(y_n, f(x_n))$ by the normalized weight w_n of the training example.

Regularized AdaBoost for two-class classification consists of the following steps (Ratsch, 2001; Ratsch et al., 1999, 2001):

$$w_n^{(1)} = 1/N \tag{5}$$

$$\varepsilon_t = \sum_{n=1}^{N} [w_n^{(t)} I(-y_n h_t(x_n))] \tag{6}$$

$$\gamma_t = \sum_{n=1}^{N} [w_n^{(t)} y_n h_t(x_n)] \tag{7}$$

$$\alpha_t = \frac{1}{2}\log\frac{1+\gamma_t}{1-\gamma_t} - \frac{1}{2}\log\frac{1+C}{1-C} \tag{8}$$

$$w_n^{(t+1)} = w_n^{(t)} \exp[-\alpha_t y_n h_t(x_n)]/Z_t \tag{9}$$

$$f(x) = \frac{1}{\Sigma_{t=1}^{T}\alpha_t}\sum_{t=1}^{T} \alpha_t h_t(x) \tag{10}$$

Here N is a number of training data points, x_n a model/classifier input set of the nth data point, y_n the corresponding class label (i.e., -1 or $+1$), T the number of boosting iterations, $\omega_n^{(t)}$ a weight of the nth data point at tth iteration, Z_t the weight normalization constant at tth iteration, $h_t(x_n) \rightarrow [-1;+1]$ the best base hypothesis (model) at tth iteration, C a regularization (soft margin) parameter, and $f(x)$ is a final weighted linear combination of the base hypotheses (models).

Boosting starts with equal and normalized weights for all training data (step (5)). A base classifier (model) $h_t(x)$ is trained using weighted error function ε_t (step (6)). If a pool of several types of base classifiers is used, then each of them is trained and the best one (according to ε_t) is chosen at the current iteration. The training data weights for the next iteration are computed in steps (7)–(9). It is clear from step (9) that at each boosting iteration, data points misclassified by the current best hypothesis (i.e., $y_n h_t(x_n) < 0$) are penalized by the weight increase for the next iteration. In subsequent iterations, AdaBoost constructs progressively more difficult learning problems that are focused on hard-to-classify patterns. This process is controlled by the weighted error function (6).

Steps (6)–(9) are repeated at each iteration until stop criteria $\gamma_t \leq C$ (i.e.,$\varepsilon_t \geq \frac{1}{2}(1 - C)$) or $\gamma_t = 1$ (i.e., $\varepsilon_t = 0$) occurs. The first stop condition means that, for the current classification problem, a classifier with accuracy better than random guess (corrected by the regularization multiplier $(1 - C)$) cannot be found. The second stop condition means that the perfect classifier is found and formulation of the next iteration classification problem focusing on misclassified examples is not needed.

Step (10) represents the final combined (boosted) model that is ready to use. The model classifies an unknown example as class $+1$ when $f(x) > 0$ and as -1 otherwise. It should be noted that the performance of the final ensemble (10) is evaluated according to the original error function given by (3) and (4) (not by (6)). Details of the more general versions of the regularized AdaBoost and its extensions are given in (Ratsch, 2001).

The original AdaBoost algorithm (Freund & Schapire, 1997) can be recovered from (5)–(10) for $C = 0$. Similar to SVM, regularization parameter C represents the so-called soft margin. Soft margin is required to accommodate the cases where it is not possible to find a boundary that fully separates data points from different classes. Regularization (soft margin) is especially important for financial applications where large noise-to-signal ratio is a typical case.

One of the most important properties of the AdaBoost is its fast convergence to a hypothesis, which is consistent with the training sample, if the base learning algorithm produces hypotheses with error rates consistently

smaller than 1/2. A theorem on the exponential convergence of the original AdaBoost ($C = 0$) states that the error (given by (3) and (4)) of the ensemble (10) on the training set is bounded above by $2^T \Pi_{t=1}^{T} \sqrt{\varepsilon_t(1-\varepsilon_t)}$ (Freund & Schapire, 1997). Thus, if the error rate in each iteration is bounded from above by $\varepsilon_t \leq \frac{1}{2} - \frac{1}{2}\mu$ (for some $\mu > 0$), then the training error ε decreases exponentially in the number of iterations: $\varepsilon \leq \exp(-T\mu^2/2)$. This and the other theorems for the original AdaBoost have been generalized for the regularized case ($C > 0$) (Ratsch, 2001).

Motivated by boosting success in classification setting, a number of researchers attempted to transfer boosting techniques to regression problems. A number of different boosting-like algorithms for regression have been proposed (e.g., Drucker, 1997; Friedman, 1999; Ratsch, 2001). However, for clarity and owing to the existence of many practically important classification problems, I focus on boosting for classification in this article.

4. BOOSTING FRAMEWORKS IN FINANCIAL AND ECONOMETRIC APPLICATIONS

Several conceptually different boosting frameworks can be proposed for financial and econometric applications. In the most straightforward way, boosting can be employed as yet another advanced data-driven algorithm expanding collection of tools used for building empirical models. However, models, built in this way, would still have "black box" nature and potentially limited stability.

In our opinion, the most appealing feature of boosting, especially, for financial applications, is that it combines the power of the advanced learning algorithm with the ability to use existing models specific to the field of interest. This important feature has never been addressed in machine learning and related literature on boosting.

The mentioned combination can be easily achieved by using well-understood and accepted models as base models (hypotheses) in a boosting framework. This allows integrating a priori problem-specific knowledge in a natural way that could lead to enhancement of the final model performance and stability, especially, in the cases of severe incompleteness of the training data, Moreover, since the final model is a weighted linear combination of the industry-standard components, the conclusions of such a model will be better accepted by practitioners and decision makers compared to a pure "black-box" model. In the following, only examples of this type of boosting applications will be discussed.

Any mature field of quantitative finance has its own set of established and well-understood models. In many cases it is very desirable to improve accuracy of the models without losing their clarity and stability. In the following, I will give several examples from different fields. One of them (symbolic volatility forecasting) will be used as an illustration of the realistic boosting application in the next section. Encouraging results and all steps of boosting application to this practically important problem will be presented.

4.1. Typical Classification Problems

It would be convenient to distinguish four large groups of the potential boosting frameworks in quantitative finance and financial econometrics. The first group includes well-defined classification models that can be used as base hypotheses in a standard boosting framework ((5)–(10)).

For example, bankruptcy prediction models based on accounting and market data are used to classify companies as bankrupt/non-bankrupt (usually for the time horizon of 1 year). These models are based on linear discriminant analysis (LDA), logit, and other statistical and machine learning frameworks (e.g., Hillegeist, Keating, Cram, & Lundstedt, 2004; Fan & Palaniswami, 2000; Fanning & Cogger, 1994). Typical industry-standard models include Altman's Z-score (LDA) (Altman, 1968), Ohlson model (logit) (Ohlson, 1980), and variations of the Black–Scholes–Merton model (Hillegeist et al., 2004). One can use the same inputs (financial ratios) and framework (LDA, logit, etc.) as in well-accepted models and build classifier on a weighted training set at every boosting iteration. Obtained boosted combination will still be based on a trusted and simple framework while it may have significantly better performance compared to the single base model.

4.2. Symbolic Time Series Forecasting

Prediction of the financial time series (stocks, market indexes, foreign exchange rates, etc.) and their volatilities is one of the most challenging and generally unresolved problems. Time series predictor with sufficient accuracy is a desired component in both trading and risk management applications. In many practical cases forecasting of the actual future value of the time series is not required. Instead it is enough to forecast symbolically encoded value.

For example, the model can predict whether the future value will be supercritical or subcritical to some threshold value or just the direction of change (up or down). In many practical problems, switching from the full

regression problem (i.e., actual value prediction) to the symbolic encoding allows to increase accuracy of the prediction, since practically unimportant small fluctuations and/or noise are removed by this procedure (Gavrishchaka & Ganguli, 2001b; Tino, Schittenkopf, Dorffner, & Dockner, 2000).

Classification problem that is obtained from the original time series forecasting problem by symbolic encoding is a second group suited for the standard boosting framework (5)–(10). Boosting can combine existing analytical and other parsimonious time series models. For example, GARCH-type framework (Engle, 1982; Bollerslev, 1986) is an industry standard for the volatility modeling. GARCH-type models can serve as base hypotheses pool in a boosting framework for the volatility forecasting. Details and preliminary results of the boosting application to the symbolic volatility forecasting will be discussed in the next section.

4.3. Portfolio Strategy Discovery and Optimization

Discovery and optimization of the portfolio strategy is an important area where boosting may also prove to be useful. Simple trading rules based on the well-known market indicators often have significant performance and stability limitations. They can also be used by many market participants easily making market efficient to these sets of trading strategies. On the other hand, adaptive strategies, capable of utilizing more subtle market inefficiencies, are often of the "black-box" type with limited interpretability and stability. Combining basic (intuitive) trading strategies in a boosting framework may noticeably increase return (decrease risk) while preserving clarity of the original building blocks.

Straightforward application of the standard boosting framework (5)–(10) for the trading strategy optimization is not possible, since it is a direct optimization and not classification problem. However, for a large class of optimization objective functions, boosting for classification can be easily transformed into a novel framework that can be called "boosting for optimization."

For example, one can require returns (r) generated by the strategy on a given time horizon (τ) to be above certain threshold (r_c). By calculating strategy returns on a series of shifted intervals of length τ and using two-class encoding ($r >= r_c$ and $r < r_c$), one obtains symbolically encoded time series as previously discussed. Encoding can be based on a more sophisticated condition that combines different measures of profit maximization and risk minimization according to the utility function of interest.

Contrary to symbolic time series forecasting discussed earlier, here the purpose is not correct classification, but rather maximization of one class

(i.e., $r >= r_c$) population. Formally, one still has classification problem where the boosting framework (5)–(10) can be applied. However, in this case, one has very uneven sample distribution between two classes. Therefore, not all theoretical results that have been rigorously proven for boosting could be automatically assumed to be valid here. Nevertheless, our preliminary empirical studies indicate consistent stability and robustness of the boosting for optimization framework when applied to technical stock/index trading.

In the case of trading strategy optimization, the usage of boosting output is also different. Instead of using weighted linear combination of the base hypotheses as a model for classification, one should use boosting weights for strategy portfolio allocation. It means that initial capital should be distributed among different base strategies in amounts according to weights obtained from the boosting. Details of boosting application to the portfolio strategy optimization will be given in our future work.

4.4. Regression Problems

Finally, the fourth group of applications explicitly requires boosting for regression (Ratsch, 2001). For example, time series forecasting models can be combined in their original form without casting into classification problem through symbolic encoding.

Another practically important application of the boosting for regression may be construction of the pricing surface for complex derivative instruments using simplified analytical or numerical valuation models as base hypotheses. This may significantly improve representation accuracy of the real market pricing surface compared to a single pricing model. Boosting-based pricing framework could be especially useful for less common or novel derivative instruments where valuation models used by different market participants are not yet standardized. This type of boosting application can be considered as a hybrid derivative pricing contrary to pure empirical pricing approaches based on machine learning algorithms (e.g., Hutchinson, Lo, & Poggio, 1994; Gencay & Qi, 2001).

5. SYMBOLIC VOLATILITY FORECASTING

In this section, I consider boosting-based framework for symbolic volatility forecasting together with all the details of its implementation and application. Stock market index and foreign exchange rate volatility are very

important quantities for derivative pricing, portfolio risk management, and as one of the components used for decision making in trading systems. It may be the main component for the quantitative volatility trading (e.g., Dunis & Huang, 2002). I start with a short overview of the existing deterministic volatility models and their limitations. Stochastic volatility models are not considered here.

A common example of the deterministic volatility models is autoregressive conditional heteroskedastic (ARCH)-type models (Engle, 1982; Bollerslev, 1986). These models assume a particular stochastic process for the returns and a simple functional form for the volatility. Volatility in these models is unobservable (latent) variable. The most widely used model of this family is generalized ARCH (GARCH) process (Bollerslev, 1986). GARCR(p,q) process defines volatility as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 \tag{11}$$

where return process is defined as

$$r_t = \sigma_t \zeta_t \tag{12}$$

Here ζ_t is an identically and independently distributed (i.i.d.) random variable with zero mean and variance 1. The most common choice for the return stochastic model (ζ_t) is a Gaussian (Wiener) process. Parameters α_i and β_i from equation are estimated from historical data by maximizing the likelihood function (LF) which depends on the assumed return distribution.

To resolve some of the GARCH limitations, a number of extensions have been proposed and used by practitioners. For example, since GARCH model depends only on the absolute values of returns (r_i^2), it does not cover leverage effect. Different forms of leverage effect have been introduced in EGARCH (Nelson, 1991), TGARCH (Zakoian, 1994), and PGARCH (Ding, Granger, & Engle, 1993) models that are given below:

$$\ln(\sigma_t^2) = \alpha_0 + \sum_{i=1}^{p} \alpha_i \frac{|r_{t-i}| + \gamma_i r_{t-i}}{\sigma_{t-i}} + \sum_{i=1}^{q} \beta_i \ln(\sigma_{t-i}^2) \tag{13}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{i=1}^{p} \gamma_i S_{t-i} r_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 \tag{14}$$

$$\sigma_t^d = \alpha_0 + \sum_{i=1}^{p} \alpha_i (|r_{t-i}| + \gamma_i r_{t-i})^d + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^d \tag{15}$$

Here γ_i is a leverage parameter. For TGARCH model, $S_{t-i} = 1$ for $r_{t-i} < 0$ and $S_{t-i} = 0$ otherwise. For PGARCH, d is a positive number that can also be estimated together with α_i, β_i, and γ_i coefficients.

GARCH and its extensions are the most common choices of the volatility model by practitioners. GARCH process can reproduce a number of known stylized volatility facts, including clustering and mean reversion (e.g., Engle & Patton, 2001; Tsay, 2002). Explicit specification of the stochastic process and simplified (linear) functional form for the volatility allows to do simple analysis of the model properties and its asymptotic behavior.

However, assumptions of the ARCH-type models also impose significant limitations. Model parameter calculation from the market data is practical only for low-order models (small p and q), i.e., in general, it is difficult to capture direct long memory effects. Volatility multiscale effects are also not covered (Dacorogna, Gencay, Muller, Olsen, & Pictet, 2001). Finally, the model gives unobservable quantity that makes it difficult to quantify the prediction accuracy and comparison with other models. Some of these restrictions are relaxed in the GARCH model extensions. However, majority of the limitations mentioned cannot be resolved in a self-consistent fashion.

All advantages and limitations of the GARCH-type models, mentioned in the previous two paragraphs, suggest that the framework accepted by many practitioners is theoretically sound and clear. It can reproduce all major signatures empirically observed in real time series. On the other hand, its forecasting accuracy may be insufficient for a number of important applications and regimes (e.g., Gavrishchaka & Ganguli, 2003).

Accuracy can be improved by using more complex models that are often of the ''black-box'' type. For example, we have recently proposed the SVM-based volatility model (Gavrishchaka & Ganguli, 2003) with encouraging results for both foreign exchange and stock market data. Previously, different NN-based frameworks have also been proposed for the same purpose (Donaldson & Kamstra, 1997; Schittenkopf, Dorffner, & Dockner, 1998). However, limited stability and explanation facility of these ''black-box'' frameworks may restrict their usage only as complementary models (Gavrishchaka & Ganguli, 2003). Boosting offers an alternative way that allows to combine all the advantages of the accepted industry models such as GARCH with significant accuracy improvement.

In many practical applications, symbolic volatility forecasting may be acceptable and even desirable due to effective filtering of noise and unimportant small fluctuations. For example, for many trading and risk management purposes, it is important (and sufficient) to know whether the future volatility value will be above or below a certain threshold. This

problem is of classification type and standard boosting framework (l)–(5) can be applied. Below I will give a detailed example of the boosting application to this problem.

I consider boosting-based framework for symbolic (threshold) volatility forecasting with a base hypotheses pool consisting of the GARCH-like models (Eqs. (11)–(15)) with different types and input dimensions. In the following, I will restrict the pool to 100 models that are obtained from all possible combinations of the model types (GARCH, EGARCH, TGARCH, and PGARCH), return input dimension (from 1 to 5), and σ input dimension (from 1 to 5). In our case, the number of the possible base hypothesis on each boosting iteration is countable and fixed. Therefore, the best model at each iteration is just one of the 100 models with a minimal classification error on a weighted training set. It is different from the more conventional usage of boosting where the available base models are not countable. Instead, at each iteration, the best model is obtained by training base models (NN, classification tree, etc.) on a weighted training set (Ratsch, 2001).

As a proxy for realized volatility that will be compared with a given threshold to give supercritical/subcritical classification (encoded as "+1" and "−1"), standard deviation of the daily returns computed on the last five business days will be used. Threshold value of the daily realized volatility is taken to be 2% that corresponds to 32% of the annualized volatility. Although the exact threshold value is problem dependent, the threshold considered here is a reasonable boundary between low volatility regime (usually taken to be 20–25% of annual volatility) and higher volatility regimes.

It is difficult to find an adequate measure of the realized volatility that can be used to test GARCH model performance (e.g., Tsay, 2002). The reason is that GARCH model outputs instantaneous (i.e., "true") volatility which is a latent (non-observable) variable. Any volatility measure computed on a real data set will be an approximation due to volatility time dependence. In many academic studies, performance of the GARCH-type models is measured by comparing model output σ_i^2 with a single day squared return r_i^2 through autocorrelation of r_i^2/σ_i^2 time series (Tsay, 2002).

However, although usage of r_i^2 is appealing due to its direct theoretical meaning (r_i^2/σ_i^2 should approach i.i.d. according to (12)), single day r_i^2 is never used as a volatility measure in practice. Daily r_i^2 time series is usually extremely noisy and can have large magnitudes that are unattainable by any practical volatility model. Volatility measure, based on a five-day standard deviation, significantly removes the noise, makes magnitude comparable to GARCH outputs, while still preserves instantaneous nature to some extent. On the other hand, standard deviation based on large data interval gives just

unconditional averaged volatility that cannot be compared with GARCH conditional volatility.

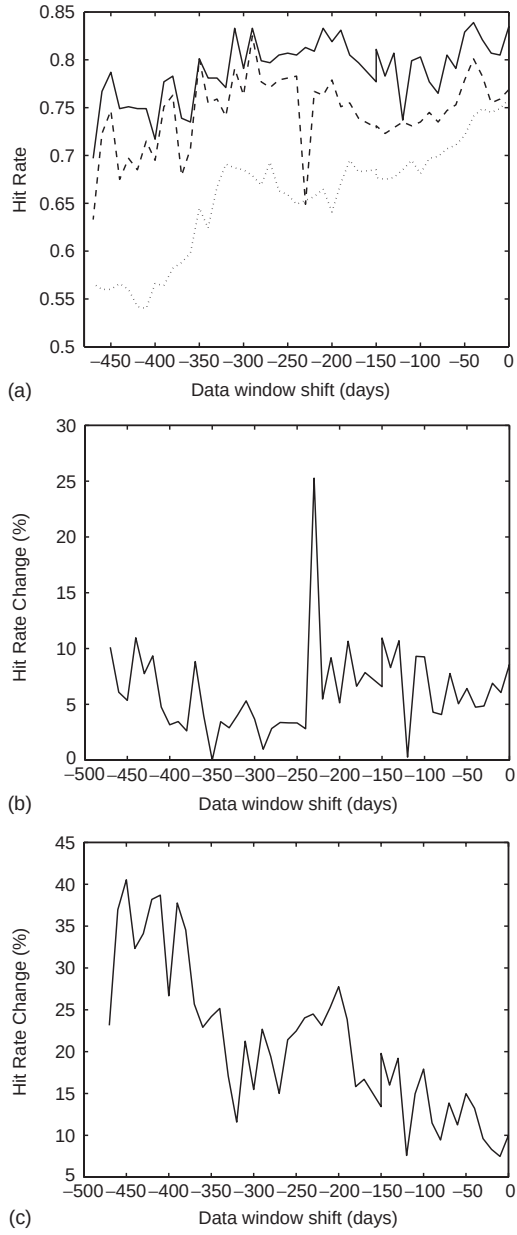
The choice of the realized volatility measure should be imposed by the requirements of the particular application. The purpose of boosting is not a "pure" test of the GARCH framework as a model for the "true" (latent) volatility. Instead, GARCH models are used as base experts that can capture the main generic features of the volatility, and boosting is used to make accurate forecasting of a particular practically important measure that is related but not identical to the "true" volatility.

I have applied boosting framework for symbolic volatility forecasting to last several years of IBM stock data. The daily closing prices adjusted for splits and dividends from finance.yahoo.com have been used. For benchmark purposes, boosting results are compared to the best model in the base hypothesis collection (as per training set) and industry-standard GARCH(1,1) model. To demonstrate result sensitivity to market conditions, data samples shifted in time are used.

The first data sample, corresponding to shift $= 0$ in all figures, includes 3 years of IBM stock daily return data (April 18, 2001–March 4, 2004). Return is computed as $r_i = \ln(S_i/S_{i-1})$, where S_i is stock closing price at day i. First 2/3 of the data sample are used as training data and the remaining part as test data. All subsequent samples are obtained by 10 business days shifting backward in time. For simplicity and clarity, I do not use validation set that allows to choose best regularization parameters C in the regularized AdaBoost algorithm (5)–(10). Instead, I use a fixed value $C = 0.05$ that shows good performance across all samples. Adaptive choice of C should only improve the performance of the boosting-based model.

In Fig. 1a, I compare performance of the boosting model (solid line), single best model from the base model collection (dashed line), and GARCH(1, 1) model (dotted line) on the training sets of all shifted data intervals (samples). As a performance measure, I use hit rate defined as a number of correctly classified data points to the total number of data points (i.e., maximum possible hit rate is 1). This measure is directly related to the

Fig. 1. (a) Hit Rate of the Boosting Model (Solid Line), Single Best Model from the Base Model Collection (Dashed Line), and GARCH(1,1) Model (Dotted Line) on the Training Sets of All Shifted Data Intervals, (b) Percentage Change of the Boosting Model Hit Rate Relative to that of the Best Single Model (Training Sets), and (c) Percentage Change of the Boosting Model Hit Rate Relative to that of the GARCH(1,1) Model (Training Sets).


0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Hit Rate
−450 −400 −350 −300 −250 −200 −150 −100 −50 0
Data window shift (days)
(a)
30
25
20
15
10
5
0
Hit Rate Change (%)
−500 −450 −400 −350 −300 −250 −200 −150 −100 −50 0
Data window shift (days)
(b)
45
40
35
30
25
20
15
10
5
Hit Rate Change (%)
−500 −450 −400 −350 −300 −250 −200 −150 −100 −50 0
Data window shift (days)
(c)

standard classification error function given by (3) and (4). It is clear that boosting models produce maximum hit rates at all times. For some samples, boosting performance is significantly superior compared to the other two models. Note that the best single model (i.e., its type and input dimensions) changes from sample to sample. The best model is the model with a maximum hit rate for a given sample.

To further quantify boosting model superiority over the two benchmark models, I compute the percentage change of the boosting model hit rate relative to the benchmark model. This number is given by $100 \times (HR - HR_0)/HR_0$, where HR and HR_0 are hit rates of the boosting and benchmark models, respectively. In Fig. 1b, this relative measure is shown for the best single model as a benchmark. One can see that in some regimes, boosting can enhance performance by up to 25%. Similar result for GARCH(1,1) model as a benchmark is shown in Fig. 1c. In this case, boosting can increase hit rate by up to 40%.

Although boosting demonstrated a superior performance on the training set, the true measure of the model performance is on the test (out-of-samplc) data. This demonstrates the generalization abilities of the model. In Fig. 2, I show exactly the same measures as in Fig. 1 but for the test sets of the shifted samples. The best single model is still the best model on the training set as would be the choice in a real application where performance on the test set is not known at the time of training. From Fig. 2, it is clear that the boosting model remains superior to both benchmark models on the test set. Moreover, the hit rate relative change even increases on the test set compared to the training set. Thus, boosting demonstrates robust out-of-sample performance and confirms its solid generalization ability.

Finally, I would like to illustrate what resources have been provided by the pool of the base models (hypotheses) such that boosting was able to achieve a significant performance improvement compared to the single base model. Figure 3 shows the number of the models that correctly classify a data point normalized to the number of all models in the pool. As an example, I consider just one sample with shift = 0 (a – training set, b – test set). It is clear that for all data points, there is a significant number of models that correctly classify this data point. However, if one considers even the best single model, a significant number of points will be misclassified. Fortunately, majority of points, misclassified by the best model, can be correctly classified by some complementary models (that may be noticeably inferior to the best model on the whole set). Thus, boosting robustly finds these complementary models at each iteration that results in a much more accurate final combined model.

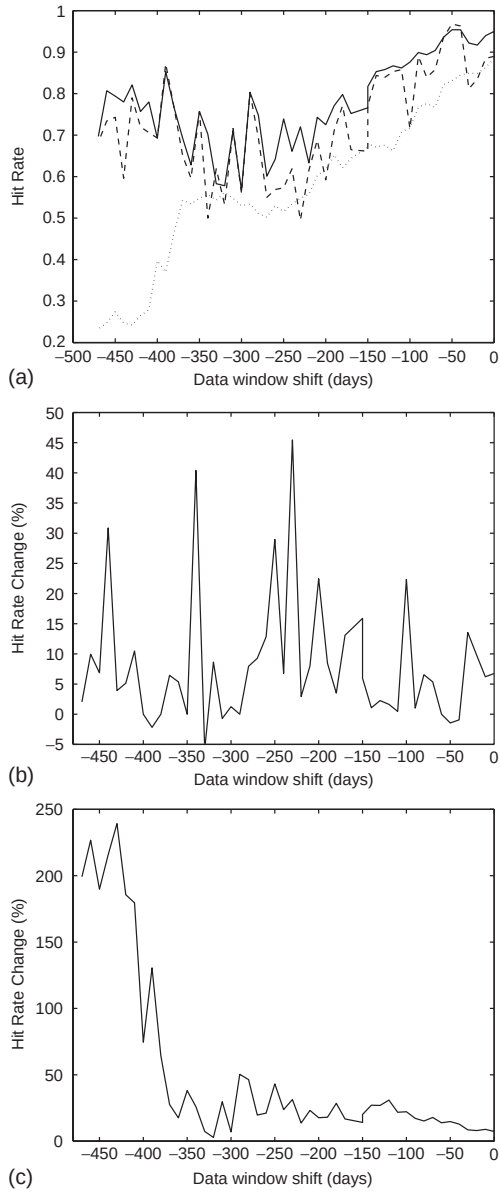


Fig. 2. The Same Measures as in Fig. 1, but for the Test (Out-of-Sample) Sets.

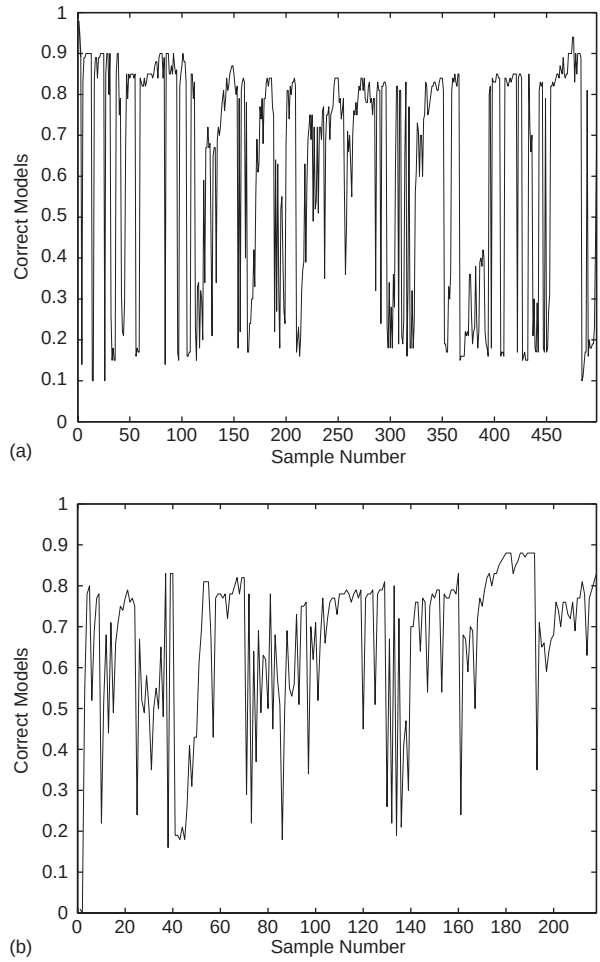


Fig. 3. Number of the Models that Correctly Classify a Data Point Normalized to the Number of all Models in the Pool for One Sample with Shift $= 0$ (a – Training Set, b – Test Set).

In the presented illustrative application, I have considered rather a restrictive pool of base volatility models. However, even with that pool, boosting was able to achieve significant improvement in performance. In practice, one can expand this pool by including more advanced and flexible models such as heterogeneous ARCH (HARCH) (Dacorogna et al., 2001)

(accounts for multiscale effects), fractionally integrated GARCH (FIGARCH) (Baillie, Bollerslev, & Mikkelson, 1996) (models short and long-term memory), and other models, including those outside the GARCH family. In addition, hybrid volatility models (e.g., Day & Lewis, 1992), that combine delayed historical inputs with available implied volatility data, could also be included in the base model pool. More detailed study of the boosting-based models for volatility forecasting is warranted.

6. DISCUSSION AND CONCLUSION

Limitations of the existing modeling frameworks in quantitative finance and econometrics, such as low accuracy of the simplified analytical models and insufficient interpretability and stability of the best data-driven algorithms have been outlined. I have made the case that boosting (a novel, ensemble learning technique) can serve as a simple and robust framework for combining the clarity and stability of the analytical models with the accuracy of the adaptive data-driven algorithms.

A standard boosting algorithm for classification (regularized AdaBoost) has been described and compared to other ensemble learning techniques. I have pointed out that boosting combines the power of the large margin classifier characterized by a robust generalization ability and open framework that allows to use simple and industry-accepted models as building blocks. This distinguishes boosting from other powerful machine learning techniques, such as NNs and SVMs, that often have ''black-box'' nature.

Potential boosting-based frameworks for different financial applications have been outlined. Some of them can use standard boosting framework without modifications. Others require reformulation of the original problem to be used with boosting. Detailed description of these frameworks and results of their application will be presented in our future works.

Details of a typical boosting operation have been demonstrated on the problem of symbolic volatility forecasting for IBM stock time series. Two-class threshold encoding of the predicted volatility has been used. It has been shown that the boosted collection of the GARCH-type models performs consistently better compared to both the best single model in the collection and the widely used GARCH(1,1) model. Depending on the regime (period in time), classification hit rate of the boosted collection on the test (out-of-sample) set can be up to 240% higher compared to GARCH(l, l) models and up to 45% higher compared to the best single model in the collection. For a training set, these numbers are 40% and 25%, respectively.

Boosting performance is also very stable. In the worst case/regime, performance of the boosted collection just becomes comparable to the best single model in the collection.

Detailed comparative studies of boosting and other ensemble learning (model combination) methods for typical econometric applications are beyond the scope of this work. However, to understand the real practical value of boosting and its limitations, such studies should be performed in the future.

ACKNOWLEDGMENTS

This work is supported by Alexandra Investment Management. I thank the referee who evaluated this paper and the editor for the valuable comments and suggestions.

REFERENCES

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*, 589–609.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H.-O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*, 3.

Barnett, J. A. (1981). Computational methods for a mathematical theory of evidence. *Proceedings of IJCAI*, *868*.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*, 451.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

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

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, *11*(7), 1493–1518.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Journal of Information Fusion*, *6*, 1.

Cai, C. Z., Wang, W. L., Sun, L. Z., & Chen, Y. Z. (2003). Protein function classification via support vector machine approach. *Mathematical Biosciences*, *185*, 111.

Chang, R. F., Wu, W.-J., Moon, W. K., Chou, Y.-H., & Chen, D.-R. (2003). Support vector machine for diagnosis of breast tumors on US images. *Academy of Radiology*, *10*, 189.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559.

Cristianini, N., & Shawe-Taylor, J. (2000). *Introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Dacorogna, M. M., Gencay, R., Muller, U., Olsen, R. B., & Pictet, O. V. (2001). *An introduction to high-frequency finance*. San Diego: Academic Press.

Day, T. E., & Lewis., C. M. (1992). Stock market volatility and the information content of stock index options. *Journal of Econometrics*, *52*, 267.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In: J. Kittler & F. Roli (Eds), *First international workshop on multiple classifier systems, Lecture Notes in Computer Science* (pp. 1–15). Heidelberg: Springer Verlag.

Ding, Z., Granger, C. W., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*, 83.

Donaldson, R. G., & Kamstra, M. (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance*, *4*, 17.

Drucker, H. (1997). Improving regressors using boosting techniques. In: D. H. Fisher (Ed.), *Proceedings of the fourteenth international conference on machine learning (ICML 1997)* (pp. 107–115). Nashville, TN, USA, July 8–12. Morgan Kaufmann, ISBN 1–55860–486–3.

Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., & Vapnik, V. (1994). Boosting and other ensemble methods. *Neural Computation*, *6*, 1289–1301.

Drucker, H., Schapire, R. E., & Simard, P. Y. (1993). Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*, 705.

Dunis, C. L., & Huang, X. (2002). Forecasting and trading currency volatility: An application of recurrent neural regression and model combination. *Journal of Forecasting*, *21*, 317.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, *50*, 987.

Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance*, *1*, 237.

Fan, A., & Palaniswami, M. (2000). Selecting bankruptcy predictors using a support vector machine approach. *Proceedings of the IEEE-INNS-ENNS international joint conference on neural networks*, *IJCNN 2000, Neural computing: New challenges and perspectives for the new millennium* (Vol. 6, pp. 354–359). Como, Italy, July 24–27.

Fanning, K., & Cogger, K. (1994). A comparative analysis of artificial neural networks using financial distress prediction. *International Journal of Intelligent Systems in Accounting, Finance, and Management*, *3*(3), 241–252.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119.

Friedman, J. H. (1999). *Greedy function approximation*. Technical report (February). Department of Statistics, Stanford University.

Gardner, A. (2004). *A novelty detection approach to seizure analysis from intracranial EEG*. Ph.D. thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology.

Gavrishchaka, V. V., & Ganguli, S. B. (2001a). Support vector machine as an efficient tool for high-dimensional data processing: Application to substorm forecasting. *Journal of Geophysical Research*, *106*, 29911.

Gavrishchaka, V. V., & Ganguli, S. B. (2001b). Optimization of the neural-network geomagnetic model for forecasting large-amplitude substorm events. *Journal of Geophysical Research*, *106*, 6247.

Gavrishchaka, V. V., & Ganguli, S. B. (2003). Volatility forecasting from multiscale and high-dimensional market data. *Neurocomputing*, *55*, 285.

Geman, S., Bienenstock, E., & Dourstat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1.

Gencay, R., & Qi, M. (2001). Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping and bagging. *IEEE Transactions on Neural Networks*, *12*(4), 726–734.

Gleisner, H., & Lundstedt, H. (2001). Auroral electrojet predictions with dynamic neural networks. *Journal of Geophysical Research*, *106*, 24, 541.

Granger, C. W. J. (1989). Combining forecasts – Twenty years later. *Journal of Forecasting*, *8*, 167.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*, 993.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, *9*, 5–34. Kluwer Academic Publishers.

Hutchinson, J. M., Lo, A., & Poggio, A. W. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance*, *31*, 851.

Iyer, R. D., Lewis, D. D., Schapire, R. E., Singer, Y., & Singhal, A. (2000). Boosting for document routing, In: *Proceedings of the ninth international conference on information and knowledge management*.

Kilian, L., & Inoue, A. (2004). Bagging time series models. *Econometric society 2004 North American summer meetings*.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Neural Information Processing Systems (NIPS)*, *7*, 231.

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

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *19*, 109.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169.

Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: An application to face detection. Conference on Computer Vision and Patterns Recognition (CVPR'97). IEEE Computer Society, 130.

Ratsch, G. (2001). *Robust boosting via convex optimization: Theory and applications*. Ph.D. thesis, Potsdam University.

Ratsch, G., et al. (1999). Regularizing AdaBoost. In: M. S. Kearns, S. A. Solla & D. A. Cohn (Eds), *Advances in neural information processing systems*, (Vol. 11, pp. 564–570). Cambridge, MA: MIT Press.

Ratsch, G., et al. (2001). Soft margins for AdaBoost. *Machine Learning*, *42*, 287.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *31*, 172.

Schapire, R. E. (1992). *The design and analysis of efficient learning algorithms*. Ph.D. thesis, Cambridge, MA: MIT Press.

Schapire, R. E., & Singer, Y. (2000). BoostTexter: A boosting-based system for text categorization. *Machine Learning*, *39*, 135.

Schapire, R. E., Freund, Y., Bartlett, P. L., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*, 1651.

Schittenkopf, C., Dorffner, G., & Dockner, E. J. (1998). Volatility prediction with mixture density networks. In: L. Niklasson, M. Boden & T. Ziemke (Eds), *ICANN '98 – Proceedings of the 8th international conference on artificial neural networks* (p. 929). Berlin: Springer.

Scholkopf, B., Tsuda, K., & Vert, J. -P. (Eds) (2004). *Kernel methods in computational biology (computational molecular biology)*. Cambridge, MA: MIT Press.

Schwenk, H., & Bengio, Y. (1997). AdaBoosting neural networks. In: W. Gerstner, A. Germond, M. Hasler & J.-D. Nicoud (Eds), *Proceedings of the Int. Conf. on Artificial Neural Networks (ICANN'97)* (Vol. 1327 of LNCS, pp. 967–972). Berlin, Springer.

Sun, X. (2002). Pitch accent prediction using ensemble machine learning. In: *Proceedings of the 7th international conference on spoken language processing (ICSLP)*. Denver, CO, USA, September 16–20.

Tino, P., Schittenkopf, C., Dorffner, G., & Dockner, E. J. (2000). A symbolic dynamics approach to volatility prediction. In: Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo & A. S. Weigend (Eds), *Computational finance 99* (pp. 137–151). Cambridge, MA: MIT Press.

Tsay, R. S. (2002). *Analysis of Financial Time Series*. New York: Wiley.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*, 1134.

Van Gestel, T., Suykens, J., Baestaens, D., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., Vandewalle, J. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks*, Special Issue on Neural Networks in Financial Engineering, *12*, 809

Vapnik, V. (1995). *The nature of statistical learning theory*. Heidelberg: Springer Verlag.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Witten, I. H., & Frank, E. (2000). *Data mining*. San Francisco, CA: Morgan Kaufmann Publishers.

Xiao, R., Zhu, L., & Zhang, H. -J. (2003). Boosting chain learning for object detection. In: *Proceedings of the ninth IEEE international conference on computer vision* (*ICCV03*). Nice, France.

Zakoian, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics Control*, *18*, 931.

This page intentionally left blank

OVERLAYING TIME SCALES IN FINANCIAL VOLATILITY DATA

Eric Hillebrand

ABSTRACT

Apart from the well-known, high persistence of daily financial volatility data, there is also a short correlation structure that reverts to the mean in less than a month. We find this short correlation time scale in six different daily financial time series and use it to improve the short-term forecasts from generalized auto-regressive conditional heteroskedasticity (GARCH) models. We study different generalizations of GARCH that allow for several time scales. On our holding sample, none of the considered models can fully exploit the information contained in the short scale. Wavelet analysis shows a correlation between fluctuations on long and on short scales. Models accounting for this correlation as well as long-memory models for absolute returns appear to be promising.

1. INTRODUCTION

Generalized auto-regressive conditional heteroskedasticity (GARCH) usually indicates high persistence or slow mean reversion when applied to financial data (e.g., Engle & Bollerslev, 1986; Ding, Granger, & Engle, 1993; Engle & Patton, 2001). This finding corresponds to the visually discernable

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 153–178
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20025-7

volatility clusters that can be found in the time series of squared or absolute returns from financial objects. The presence of change-points in the data, however, prevents GARCH from reliably identifying the long time scale that is associated with the high persistence (Diebold, 1986; Lamoureux & Lastrapes, 1990; Mikosch & Starica, 2004; Hillebrand, 2005). Recently, the possibility of the presence of several correlation time scales in financial volatility data has been discussed (Fouque, Papanicolaou, Sircar, & Sølna, 2003; Chernov, Gallant, Ghysels, & Tauchen, 2003; Gallant & Tauchen, 2001).

Using six different time series of daily financial data, we show that autoregressive conditional heteroskedasticity (ARCH), GARCH, and Integrated GARCH (IGARCH) fail to capture a short correlation structure in the squared returns with a time scale of less than 20 days. This serial correlation with low persistence can be used to improve the short-term forecast error from GARCH models.

Several models that are being proposed to capture different time scales are discussed and estimated on our data set. Unfortunately, none is able to fully exploit the information contained in the short scale in our data set. In correspondence to the results of Ding and Granger (1996), ARFIMA estimation of absolute, as opposed to squared, returns yields a consistent improvement of the 1-day forecast on our holding sample. Wavelet transformation of the volatility time series from our data set reveals that there is correspondence between long- and short-correlation scales. This supports the findings of Müller et al. (1997) and Ghashgaie, Breymann, Peinke, Talkner, and Dodge (1996). The heterogeneous ARCH (HARCH) model of Müller et al. (1997), which allows for correlation between long and short-scales, indeed uses the short-scale information for the stock market part of our data set best.

In Section 2, we introduce our data set and discuss the ARCH, GARCH, and IGARCH results, which turn out to be very typical for financial data. Section 2.3 identifies the short-correlation scale that is being missed by GARCH, using spectral methods as well as ARMA estimation, which provides a benchmark for forecast improvement over GARCH. In Section 3, we segment our time series using recent results for change-point detection in volatility data. GARCH is then estimated locally on the segments. As change-points cause GARCH to show spurious high persistence, the idea is that accounting for change-points may reveal the short scale within the segments. This is indeed the case, but does not help to improve forecasts for our holding sample. Section 4 discusses and estimates an array of models designed to capture several time scales in volatility. We begin with fractionally integrated models and estimate ARFIMA and FIGARCH on our data set. ARFIMA estimation of absolute returns yields a consistent

improvement of the 1-day ahead forecast on our holding sample. Next, a two-scale version of GARCH by Engle and Lee (1999) is estimated. Section 4.3 demonstrates the correspondence between long and short scales with wavelet analysis. The HARCH model of Müller et al. (1997) is then estimated. We conclude that, on our data set, none of the considered models captures the short scale as well as the simple ARMA fit to the residual in Section 2.3. For the stock market section of our data, HARCH yields the relatively largest improvements in the short-term forecast errors.

2. INTEGRATED VOLATILITY

2.1. Data

We use an array of daily financial data. Four series measure the U.S. stock market: the S&P 500 index (sp) and the Dow Jones Industrial Average (dj), obtained from Datastream, as well as the CRSP equally weighted index (cre) and the CRSP value-weighted index (crv). Next, we consider the exchange rate of the Japanese Yen against the U.S. Dollar (yd, Datastream series BBJPYSP) and the U.S. federal funds rate (ffr). These series cover the period January 4, 1988 through Decmber 31, 2003 and contain 4,037 observations, except for the Yen/Dollar exchange rate, which is recorded in London and therefore contains only 3,991 observations.

2.2. ARCH and GARCH

In the ARCH(p) model (Engle, 1982) the log-returns of a financial asset are given by

$$r_t = m_t + \varepsilon_t, t = 1, \ldots, T, \tag{1}$$

$$\varepsilon_t | \mathscr{F}_{t-1} \sim \mathscr{N}(0, h_t), \tag{2}$$

$$h_t = \sigma^2 + \sum_{i=1}^{p} \alpha_i (\varepsilon_{t-1}^2 - \sigma^2), \tag{3}$$

where $m_t = \mathbb{E}_{t-1} r_t$ is the conditional mean function, $\mathscr{F}_t$ denotes the filtration that models the information set, and $\sigma^2 = \omega/(1 - \sum_{i=1}^{p} \alpha_i)$ is the unconditional mean with ω being a constant. For the conditional mean m_t different models are used, for example ARMA processes, regressions on exogenous

variables, or simply a constant. Since we focus on the persistence of the conditional volatility process, we will use a constant in this study.

The GARCH(p,q) model (Bollerslev, 1986) adds the term $\sum_{i=1}^{q}\beta_i(h_{t-i} - \sigma^2)$ to the conditional variance Eq. (3). The unconditional mean becomes $\sigma^2 = \omega/(1 - \sum_{i=1}^{p}\alpha_i - \sum_{i=1}^{q}\beta_i)$ Empirically, GARCH models often turn out to be more parsimonious than ARCH models since these often need high lag orders to capture the dynamics of economic time series.

There are different ways to measure persistence in the context of GARCH models, as discussed, for example, in Engle and Patton (2001) and Nelson (1990). The most commonly used measure is the sum of the autoregressive parameters,

$$\lambda := \sum_{i=1}^{p}\alpha_i + \sum_{i=1}^{q}\beta_i. \tag{4}$$

The parameter λ can be interpreted as the fraction of a shock to the volatility process that is carried forward per unit of time. Therefore $(1-\lambda)$ is the fraction of the shock that is washed out per unit of time. Hence, $1/(1-\lambda)$ is the average time needed to eliminate the influence of a shock. The closer λ is to unity, the more persistent is the effect of a change in h_t. The stationarity condition is $\lambda < 1$. A stationary but highly persistent process returns slowly to its mean, a process with low persistence reverts quickly to its mean. Different persistences therefore imply different times of mean reversion. In this study, the term "time scales" is understood in this sense.

Estimations of the conditional volatility process with GARCH models usually indicate high persistence (Engle & Bollerslev, 1986; Bollerslev, Chou, & Kroner, 1992; Bollerslev & Engle, 1993; Bollerslev, Engle, & Nelson, 1994; Baillie, Bollerslev, & Mikkelsen, 1996; Ding, Granger, & Engle, 1993, Ding & Granger, 1996; Andersen & Bollerslev, 1997). The estimated sum λ of the autoregressive parameters is often close to one. In terms of time scales, $\lambda = 0.99$ corresponds to 100 days or 5 months mean reversion time for daily data, whereas $\lambda = 0.999$ corresponds to 1000 days or 4 years. Therefore, it is difficult to identify the long scale numerically and interpret it economically.

Engle and Bollerslev (1986) suggested that the integrated GARCH, or IGARCH model reflect the empirical fact of high persistence. In the IGARCH model, the likelihood function of the GARCH model is maximized subject to the constraint that $\lambda = 1$. In the IGARCH(1, 1) specification this amounts to the parameter restriction $\beta = 1 - \alpha$. The estimated conditional volatility process is not stationary, a shock has indefinite influence on the level of volatility.

Table 1 compares GARCH, ARCH, and IGARCH estimations on our data set. Within the class of model orders up to GARCH(5,5), the Bayes Information Criterion favors GARCH(1, 1) for all series. Contrary to that, it indicates lag orders between 8 and 20 for ARCH estimations. For GARCH estimations, the sum of the autoregressive parameters is almost one for all series, with the CRSP equally weighted index showing the lowest persistence with a mean reversion time of 27 days and the volatility of the federal funds rate indicating non-stationarity.

The ARCH estimations show a very different picture of persistence, λ is of the order of 0.80–0.85 for the stock market, 0.5 for the yen–dollar exchange rate, and still close to one for the federal funds rate. This indicates that another, shorter scale of the order of a few days may be present in the data.

Table 1. ARCH, GARCH, and IGARCH Estimation.

	sp	dj	cre	crv	ffr	yd
Panel A: GARCH(1, 1) ($h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}$)						
$\alpha+\beta$	0.995	0.990	0.963	0.991	1.0	0.976
MAE(1)	7.24e-5	4.63e-5	3.36e-5	7.08e-5	0.0254	3.35e-5
MAE(20)	6.24e-5	5.33e-5	3.99e-5	5.76e-5	0.0182	3.18e-5
MAE(60)	4.59e-5	4.06e-5	3.95e-5	4.48e-5	0.0158	2.89e-5
Panel B: ARCH ($h_t = \omega + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2$)						
p	14	11	8	14	20	9
$\sum \alpha_i$	0.855	0.796	0.831	0.866	0.989	0.494
MAE(1)	1.48	1.58	1.44	1.34	0.64	0.85
MAE(20)	1.03	1.03	1.02	1.04	0.90	0.98
MAE(60)	1.03	1.02	1.00	1.02	0.91	1.03
Panel C: IGARCH(1, 1) ($h_t = \omega + \alpha\varepsilon_{t-1}^2 + (1-\alpha)h_{t-1}$)						
α	0.047	0.058	0.013	0.072	0.169	0.050
MAE(1)	1.05	1.03	1.09	1.07	0.98	0.94
MAE(20)	1.31	1.39	1.31	1.41	2.86	1.06
MAE(60)	1.64	1.63	1.39	1.74	6.77	1.15

Note: The sample consists of daily returns of the S&P 500 index (sp), Dow Jones Industrial Average (dj), CRSP equally-weighted index (cre), CRSP value-weighted index (crv), federal funds rate (ffr), and yen per dollar exchange rate (yd) from January 4, 1988 through October 6, 2003. The MAEs are calculated using the holding sample October 7, 2003 through December 31, 2003. The MAE for the ARCH and IGARCH models are stated in percentages of the forecast error from the GARCH(1, 1) model.

The last three rows of every panel show the mean absolute errors (MAE) for a 1-, 20-, and 60-day forecast of the conditional volatility process when compared to the squared returns of the holding sample October 7, 2003 through December 31, 2003. That is,

$$\begin{aligned} \text{MAE}(1) &= \left|\varepsilon_t^2 - h_t\right|, \quad t = \text{Oct 7, 2003}, \\ \text{MAE}(20) &= \tfrac{1}{20} \sum_{t=\text{Oct 7, 2003}}^{\text{Nov 3, 2003}} \left|\varepsilon_t^2 - h_t\right|, \\ \text{MAE}(60) &= \tfrac{1}{60} \sum_{t=\text{Oct 7, 2003}}^{\text{Dec 31, 2003}} \left|\varepsilon_t^2 - h_t\right|. \end{aligned}$$

The mean absolute errors of the GARCH(1, 1) forecast are taken as a benchmark. The forecast errors of other models are expressed as percentages of the GARCH(1, 1) error. In terms of forecast performance on the holding sample considered here, GARCH dominates ARCH for the stock market data. For the federal funds rate and the exchange rate, ARCH seems to be a better forecast model.

One might suspect that the stock market volatility is in fact integrated and that GARCH(1, 1) with λ close to one reflects this fact better, thereby providing better forecasts than ARCH. This interpretation, however, cannot hold as IGARCH(1, 1) forecasts are worse than GARCH across all series and forecast horizons, the only exceptions being the single-day forecasts of the exchange rate and the federal funds rate. Further, IGARCH performs worse than ARCH for the 20-day and 60-day forecasts of stock market volatility but generally better on the single-day horizon. This is a counterintuitive result as the non-stationarity of IGARCH should reflect the long-term behavior better than the short term.

Note that the GARCH(1, 1) and ARCH models in Table 1 display an unexpected pattern in the forecast error: The error often decreases with increasing forecast horizon. This indicates that these models miss short-term dynamics and capture long-term dynamics better.

2.3. The Missed Short Scale

The GARCH model aims to capture the entire correlation structure of the volatility process ${\varepsilon_t}^2$ in the conditional volatility process h_t. Therefore, we can define the residual $v_t := \varepsilon_t^2 - h_t$. This residual is white noise. From the distribution assumption (2) of the GARCH model, which can be written as

$\varepsilon_t^2 = \eta_t^2 h_t$, where η_t is a standard normal random variable, we have that

$$\mathbb{E}v_t = \mathbb{E}(\mathbb{E}_{t-1}(\varepsilon_t^2 - h_t)) = \mathbb{E}(\mathbb{E}_{t-1}(\eta_t^2 h_t - h_t)) = 0, \tag{5}$$

$$\mathbb{E}v_t v_s = \mathbb{E}((\eta_t^2 - 1)(\eta_s^2 - 1)h_t h_s) = \begin{cases} 0 & \text{for } t \neq s, \\ 2\mathbb{E}h_t^2 & \text{for } t = s. \end{cases} \tag{6}$$

The latter is shown to exist in Theorem 2 of Bollerslev (1986). Therefore, if GARCH is an accurate model for financial volatility, the estimated residual process $\hat{v}_t = \hat{\varepsilon}_t^2 - \hat{h}_t$ should not exhibit any serial correlation.

Figures 1–3 show the estimated averaged periodograms of the residuals of the GARCH estimations in Table 1. The averaged periodogram is estimated by sub-sampling with a Tukey–Hanning window of 256 points length allowing for 64 points overlap. A Lorentzian spectrum model

$$h(w) = a + b/(c^2 + w^2) \tag{7}$$

is fitted to the periodogram, where w denotes the frequencies and (a, b, c) are parameters. The average mean reversion time is obtained as a function of the parameter c.[1] Except for the CRSP equally weighted index and the federal funds rate, the series show strong serial correlation in the residual. The persistence time scale is of the order of five days.

This finding is confirmed when ARMA models are fitted to the residuals. Table 2 reports the parameter estimates and the mean absolute errors of the forecasts for the volatility process ${\varepsilon_t}^2$ when the serial correlation in the residual v_t is taken into account. The errors are stated as percentages of the errors from the simple GARCH estimations in Table 1. The first order autoregressive parameters are estimated around 0.8, corresponding to the 5-day mean reversion time scale found from the Lorentzian model and corresponding to the estimated persistence from the ARCH models in Table 1. Note that the forecast errors can be substantially reduced by correcting for this correlation, which is exogenous to the GARCH model. Using the residual ${\varepsilon_t}^2/h_t$, which is a truncated standard normal under GARCH, or adding autoregressive terms to the mean equation does not alter these findings.

We are now able to explain the counterintuitive pattern of forecast errors from the GARCH model. The fact that the short correlation scale is not captured leads to inflated errors on short forecast horizons. The correction for the short scale therefore reduces the errors at the 1-day and at the 20-day horizon much more than at the 60-day horizon.

In the following sections, we will explore different ways to account for the short scale in GARCH-type models. We will evaluate the forecast performance

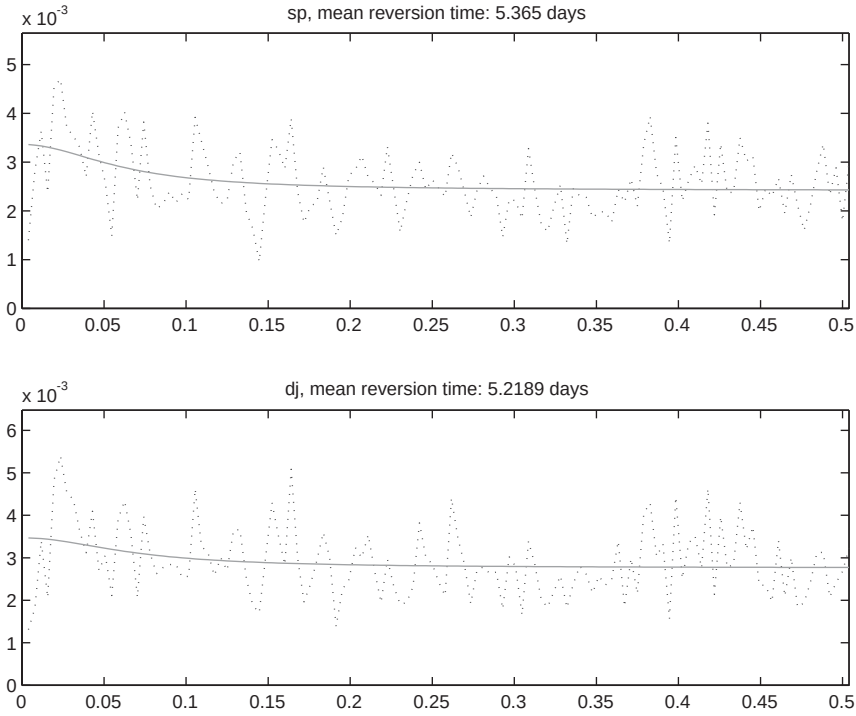


Fig. 1. Periodogram Estimation I. Estimation of the Power Spectra (Dotted Line) of the Residual Processes $\hat{v}_t = \hat{\varepsilon}_t^2 - \hat{h}_t$ of the S&P 500 and Dow Jones Series and Nonlinear Least Squares fit of a Lorentzian Spectrum (Solid Line). The Estimate of the Average Mean Reversion Time is Computed from the Lorentzian.

of these models against the benchmark of the exogenous correction for the short scale in Table 2.

3. LOCAL GARCH ESTIMATION

The apparent integrated or near-integrated behavior of financial volatility may be an artifact of changing parameter regimes in the data. Several authors have studied this phenomenon in a GARCH framework with market data and in simulations (Diebold, 1986; Lamoureux & Lastrapes, 1990; Hamilton & Susmel, 1994; Mikosch & Starica, 2004). Hillebrand (2005) shows that if a GARCH model is estimated globally on data that were

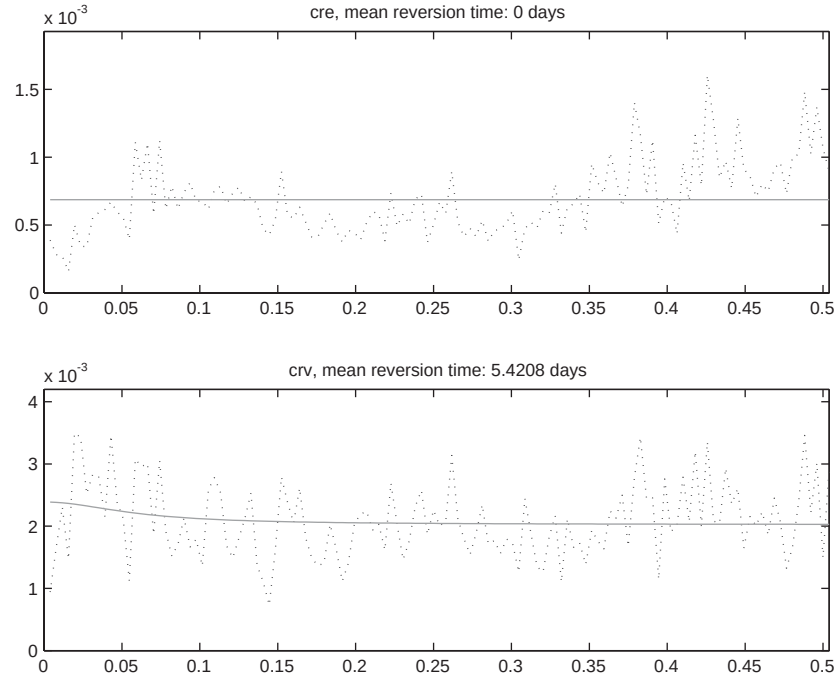


Fig. 2. Periodogram Estimation II. Estimation of the Power Spectra (Dotted Line) of the Residual Processes $\hat{v}_t = \hat{\varepsilon}_t^2 - \hat{h}_t$ of the CRSP Equally Weighted Index and the CRSP Value-Weighted Index and Nonlinear Least Squares fit of a Lorentzian Spectrum (Solid Line). The Estimate of the Average Mean Reversion Time is Computed from the Lorentzian.

generated by changing local GARCH models, then the global λ will be estimated close to one, regardless of the value of the data-generating λ within the segments. In other words, a short mean reversion time scale in GARCH plus parameter change-points exhibits statistical properties similar to a long-mean reversion time scale.

Tables 3 and 4 show the estimation of GARCH(1, 1) models on segmentations that were obtained using a change-point detector proposed by Kokoszka and Leipus (1999). The detector statistic is given by

$$U(t) = \sqrt{T}\frac{t(T-t)}{T^2}\left(\frac{1}{t}\sum_{j=1}^{t} r_j^2 - \frac{1}{T-t}\sum_{j=t+1}^{T} r_j^2\right). \tag{8}$$

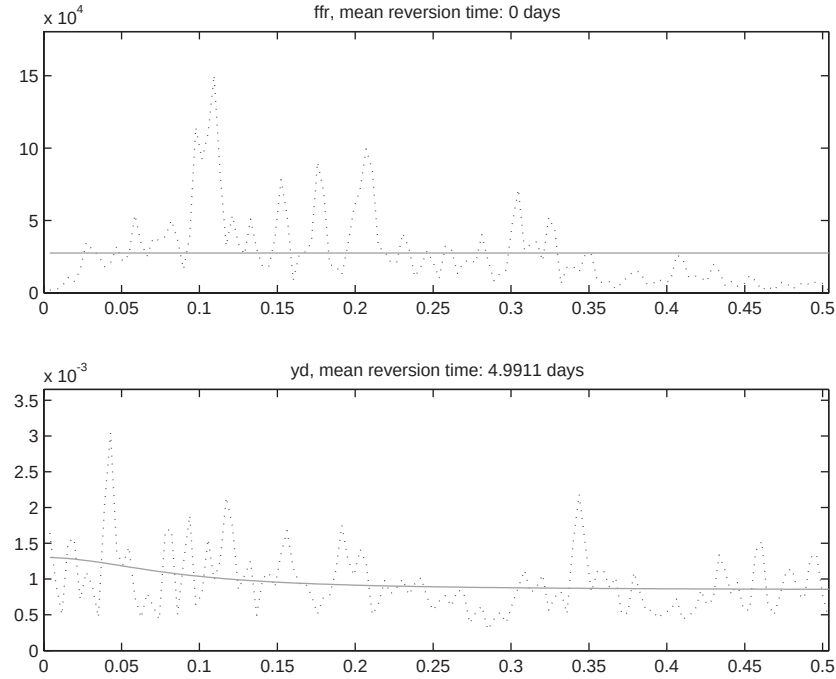


Fig. 3. Periodogram Estimation III. Estimation of the Power Spectra (Dotted Line) of the Residual Processes $\hat{v}_t = \hat{\varepsilon}_t^2 - \hat{h}_t$ of the Federal Funds Rate and Yen–Dollar Exchange Rate Series and Nonlinear Least Squares Fit of a Lorentzian Spectrum (Solid Line). The Estimate of the Average Mean Reversion Time is Computed from the Lorentzian.

The estimator of the single change-point in the sample of T observations of the squared returns ${r_t}^2$ is obtained as

$$\hat{\tau} = \min\left\{\tau : |U(\tau)| = \max_{t\in\{1,\ldots T\}}\{U(t)|\right\}. \tag{9}$$

Kokoszka and Leipus (1999) show that the statistic $U(t)/\sigma$ converges in distribution to a standard Brownian bridge, where $\sigma^2 = \sum_{j=-\infty}^{\infty}\mathrm{cov}(r_t^2, r_{t+j}^2)$. Therefore, confidence intervals can be calculated from the distribution of the maximum of a Brownian bridge. We follow Andreou and Ghysels (2002) and use the VARHAC estimator of Den Haan and Levin (1997) for σ. Change-points at a significance level of 0.05 or less are considered in Tables 3 and 4.

Table 2. Forecast Correction Using Short Scale.

	sp	dj	cre	crv	ffr	yd
$v_t = \sum_{j=1}^{p} \phi_j v_{t-j} + \sum_{j=1}^{q} \psi_j \eta_{t-j} + \eta_t$, η white noise						
Spec	(1,1)	(1,1)	(1,2)	(1,1)	(2,2)	(1,1)
ϕ_1	0.82	0.80	0.79	0.78	0.73	0.53
ϕ_1					0.08	
ψ_1	−0.78	−0.78	−0.91	−0.75	−0.48	−0.44
ψ_2			0.26		−0.44	
MAE(1)	0.10	0.44	0.42	0.19	0.20	0.41
MAE(20)	0.76	0.81	0.95	0.81	0.79	0.97
MAE(60)	0.89	0.91	0.98	0.92	0.92	0.99

Note: ARMA parameter estimates and mean absolute errors for the 1-day, 20-days, and 60-days forecast of the global GARCH(1, 1) model correcting for the short scale found in the residual $v_t = \varepsilon_t^2 - h_t$. The ARMA specification was found by including only parameters significant at the 0.05 level or lower. The errors are reported in percent of the global GARCH(1, 1) forecast error in Table 1.

Table 3. Change-point Detection.

sp			dj			cre		
Segment	Prob	λ	Segment	Prob	λ	Segment	Prob	λ
01/04/88			01/04/88			01/04/88		
12/30/91	0.00	0.925	05/16/91	0.00	0.930	10/16/97	0.00	0.849
01/08/96	0.00	0.948	10/13/92	0.01	0.517	10/06/03		0.964
12/06/96	0.01	0.909	12/15/95	0.04	0.905			
08/14/97	0.00	0.989	03/26/97	0.00	0.942			
10/06/03		0.953	10/06/03		0.960			
$\bar{\lambda}$		0.944			0.901			0.892

Note: Local GARCH(1, 1) estimations on segments identified by the change-point detector of Kokoszka and Leipus (1999).

In fact, the average estimated λ within the segments indicates lower persistence than the global estimates. The implied mean reversion time scales range from about 3 days for the yen per dollar exchange rate to about 18 days for the S&P 500 index.

However, while the approach uncovers a short scale, its forecast success is disappointing. Table 5 shows the mean absolute error from forecasts of the GARCH(1, 1) model of the last segment. Two offsetting effects are at work. While we can expect the local estimation approach to capture the short-run

Table 4. Local GARCH(1, 1) Estimations.

crv			ffr			yd		
Segment	Prob	λ	Segment	Prob	λ	Segment	Prob	λ
01/04/88			01/04/88			01/04/88		
12/30/91	0.00	0.916	09/04/90	0.02	0.991	05/07/97	0.02	0.959
12/15/95	0.00	0.900	02/22/91	0.00	0.522	09/27/99	0.00	0.902
03/26/97	0.04	0.872	01/21/93	0.00	0.933	10/06/03		0.010
10/15/97	0.00	0.937	10/06/03		0.704			
10/06/03		0.973						
$\bar{\lambda}$		0.928			0.764			0.706

Note: The segments were identified by the change-point detector of Kokoszka and Leipus (1999).

Table 5. Forecast Correction Using Local Estimations.

	sp	dj	cre	crv	ffr	yd
MAE(1)	1.56	1.64	1.36	1.41	1.06	1.06
MAE(20)	1.21	1.24	1.02	1.21	1.37	1.11
MAE(60)	1.25	1.24	0.97	1.18	1.52	1.21

Note: Mean absolute errors for the 1-, 20-, and 60-days forecast of the GARCH(1, 1) model for the last segment identified by the Kokoszka and Leipus (1999) change-point detector. The MAEs for the holding sample October 7, 2003 through December 31, 2003 are expressed in percent of the global GARCH(1, 1) forecast error as reported in Table 1.

dynamics of the series better, the shorter sample sizes increase the estimation error. In the monitoring situation, where the change-point must be detected as observations arrive, the time lag necessary to recognize a parameter change will further deteriorate the forecast performance. The MAEs of Table 5 are again stated as percentages of the global forecast errors and show that in terms of forecast success, the global approach yields better results on our holding sample.

4. MODELS OF MULTIPLE TIME SCALES

In the local estimations in Section 3 (Tables 3 and 4), we have seen that accounting for change-points and estimating GARCH on segments reduces the estimated persistence. The mean reversion time estimated from the average persistence across segments stays below 20 days as opposed to the

global estimation in Table 1, where four out of six volatility time series display a mean reversion time of more than 100 days. These findings suggest the interpretation that financial volatility displays low persistence with occasional structural breaks, which lead to apparent high persistence in estimations that ignore the change-points. Another possible interpretation is that two or more persistence time scales influence the volatility processes simultaneously and continuously in daily data. We study this interpretation in Sections 4.1 and 4.2. A third possible interpretation is that the scales remain only for some limited time and then die out and are replaced by different scales. They may also influence each other. Section 4.3 deals with this case.

A short mean reversion time scale in financial volatility has been found in many different recent studies, mostly in the context of stochastic volatility models. Gallant and Tauchen (2001) estimate a stochastic volatility model with two different volatility drivers. Estimating their model for daily returns on the Microsoft stock, one driver assumes high persistence, the other assumes low persistence. Fouque et al. (2003) and Chernov et al. (2003) propose and discuss multi-scale stochastic volatility models. LeBaron (2001) simulates a stochastic volatility model where the volatility process is an aggregate of three different persistence time scales and shows that the simulated time series display long-memory properties. This corresponds to Granger's (1980) finding that aggregation over several scales implies long-memory behavior. These findings suggest that both, long and short scale act continuously and simultaneously in the data.

4.1. Fractional Integration

Assume that the time series x_t, when differenced d times, results in a time series y_t that has an ARMA representation. Then, the time series x_t is called integrated of order d, denoted $x_t \sim I(d)$. Fractionally integrated time series, where d is not an integer, are a discretized version of fractional Brownian motion.[2] Granger (1980) and Granger and Joyeux (1980) show that the autocorrelation function of a fractionally integrated x_t is given by

$$\rho_k = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(k+1-d)} \equiv A_d k^{2d-1},$$

for $d \in (0, 1/2)$, where A_d is a constant. The sum over the autocorrelations does not converge, so that it is a suitable model for long memory. If $d \in (1/2, 1)$, the process has infinite variance and is therefore non-stationary but still

mean reverting. If $d>1$, the process is non-stationary and not mean reverting. For $d\in(-1/2, 0)$, the process is anti-persistent.

Granger (1980) shows that aggregation of a large number of processes with different short mean reversion times yields a fractionally integrated process. In particular, he shows that the process

$$x_t = \sum_{j=1}^{N} y_{j,t},$$

where the $y_{j,t}$ are AR(1) models

$$y_{j,t} = \alpha_j y_{j,t-1} + \varepsilon_{j,t}, \quad j = 1, \ldots, N,$$

with $\varepsilon_{j,t}$ independent white noise processes, is a fractionally integrated process with order d depending on the statistical properties of the α_j. This is particularly appealing for economic time series, which are often aggregates. The surprising result is that an aggregate of many short scales can exhibit long memory properties.

Geweke and Porter-Hudak (1983) provide the standard estimation method for d. The estimator is obtained as the negative of the estimate $\hat{b}$ of the slope in the regression

$$\log I(w_j) = a + b \log(4 \sin^2(w_j/2) + \varepsilon_j,$$

where $I(w_j)$ is the periodogram at the harmonic frequencies $w_j = j\pi/T$ and T is the sample size.

Most financial time series exhibit significant long memory in the sense of the Geweke and Porter-Hudak (1983) estimator. Table 6 shows the estimated d for the squared and the absolute returns of the time series considered here. The estimates are significantly larger than zero and below 1/2, indicating stationary long memory. The only exception are the absolute returns of the CRSP equally weighted index, which may be non-stationary but still mean reverting. These findings may indicate the presence of multiple overlaying time scales in the data in the sense of Granger (1980).

For the absolute returns calculated from the time series considered here, we estimate an ARFIMA model using Ox (Doornik, 2002) and the ARFIMA package for Ox (Doornik & Ooms, 1999). We choose to model absolute returns since they exhibit an even richer correlation structure than squared returns (Ding & Granger, 1996) and a model for absolute returns allows to forecast squared returns as well. The ARFIMA(p,d,q) model for the time series y_t is defined as

$$\Phi(L)(1-L)^d y_t = \Psi(L)\eta_t, \tag{10}$$

Table 6. Estimation of Fractional Integration.

	sp	dj	cre	crv	ffr	yd
Panel A: r_t^2						
d	0.354	0.328	0.390	0.387	0.238	0.319
	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)
Panel B: $\lvert r_t^2 \rvert$						
d	0.418	0.428	0.507	0.436	0.343	0.422
	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)

Note: Results of the Geweke and Porter-Hudak (1983) test for the daily S&P 500 index (sp), Dow Jones Industrial Average (dj), CRSP equally weighted index (cre), CRSP value-weighted index (crv), federal funds rate (ffr), and yen per dollar exchange rate (yd) from January 4, 1988 through December 30, 2003. The null hypothesis is $d = 0$, the alternative $d \neq 0$. All estimates are significant according to all common significance levels. The smoothed periodogram was estimated using a 1,024 points window with 256 points overlap. The $g(T)$ function in Theorem 2 of Geweke and Porter-Hudak (1983) was set to $[T0.55] = 96$. The numbers in parentheses are asymptotic standard errors.

where L is the lag operator, η_t is white noise with mean zero and variance σ^2,

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p,$$

is the autoregressive lag polynomial, and

$$\Psi(L) = 1 + \psi_1 L + \psi_2 L^2 + \cdots + \psi_p L^p,$$

is the moving-average lag polynomial. We assume that all roots of the lag polynomials are outside the unit circle. The fractional differencing operator $(1-L)^d$, where $d \in (-0.5, 0.5)$, is defined by the expansion

$$(1 - L)^d = 1 - dL + \frac{d(d-1)}{2!} L^2 - \frac{d\,(d-1)(d-2)}{3!} L^3 + \cdots \qquad (11)$$

In estimations, a truncation lag has to be defined; we use 100 in this study.

We determine the ARFIMA lag orders p and q by significance according to t-values. Table 7 reports the estimation results and the forecast errors in percent of the GARCH(1, 1) forecast error. The estimates of the fractional integration parameter d are highly significant across all time series. On the 1-day horizon the forecast error is smaller than the forecast error of the GARCH(1, 1) model across all time series. Except for the yen per dollar exchange rate, the forecast performance on the holding sample is worse than GARCH(1, 1) for the 20- and 60-day forecast horizons.

Table 7. ARFIMA Estimation.

	sp	dj	cre	crv	ffr	yd
$(1-\phi L)(1-L)^d\lvert r_t\rvert = \sum_{j=0}^{q}\psi_j\eta_{t-j}+\eta_t$, η white noise						
spec	(1,d,1)	(1,d,1)	(1,d,0)	(1,d,1)	(1,d,2)	(1,d,1)
d	0.407	0.400	0.404	0.404	0.415	0.300
ϕ	0.228	0.247	−0.244	0.153	0.558	0.343
ψ_1	−0.605	−0.616		−0.537	−0.515	−0.567
ψ_2					−0.244	
MAE(1)	0.81	0.64	0.62	0.79	0.86	0.43
MAE(20)	1.06	1.11	1.04	1.10	2.38	0.91
MAE(60)	1.31	1.33	1.05	1.32	3.48	0.99

Note: The sample consists of absolute returns of the daily S&P 500 index (sp), Dow Jones Industrial Average (dj), CRSP equally weighted index (cre), CRSP value-weighted index (crv), federal funds rate (ffr), and yen per dollar exchange rate (yd) from January 4, 1988 through October 6, 2003. All parameter estimates are significant according to all common confidence levels. The MAEs for the holding sample October 7, 2003 through December 31, 2003 are expressed in percent of the GARCH(1, 1) error as reported in Table 1.

Baillie et al. (1996) proposed a GARCH model that allows for a fractionally integrated volatility process. The conditional variance Eq. (3) can be rewritten as an ARMA process for ${\varepsilon_t}^2$

$$[(1-\alpha(L)-\beta(L)]\varepsilon_t^2 = \omega + [1-\beta(L)]v_t, \tag{12}$$

where $v_t = \varepsilon_t^2 - h_t$. The coefficients $\alpha(L) = \alpha_1 L + \cdots + \alpha_p L^p$ and $\beta(L) = \beta_1 L + \cdots + \beta_q L^q$ are polynomials in the lag operator L. The order of the polynomial $[(1-\alpha(L)-\beta(L)]$ is $\max\{p, q\}$.

The residual v_t is shown to be white noise in (5) and thus, (12) is indeed an ARMA process. If the process ${\varepsilon_t}^2$ is integrated, the polynomial $[(1-\alpha(L)-\beta(L)]$ has a unit root and can be written as $\Phi(L)(1-L)$ where $\Phi(L) = [(1-\alpha(L)-\beta(L)](1-L)^{-1}$ is of order $\max\{p, q\}-1$. Motivated by these considerations the fractionally integrated GARCH, or FIGARCH model defines the conditional variance as

$$\Phi(L)(1-L)^{\delta}\varepsilon_t^2 = \omega + [1-\beta(L)]v_t, \tag{13}$$

where $\delta \in (0,1)$.

Table 8 reports the parameter estimates and MAEs of the forecast from a FIGARCH(1, δ, 1) model. The estimation was carried out using the GARCH 2.2 package of Laurent and Peters (2002) for Ox.

Table 8. FIGARCH Estimation.

	sp	dj	cre	crv	ffr	yd
$(1-(\alpha+\beta)L)\,(1-L)^{-1}(1-L)^{\delta}\,\varepsilon_t^2=\omega+(1-\beta L)v_t,\ v_t=\varepsilon_t^2-h_t.$						
δ	0.428	0.417	0.708	0.404	0.289	0.251
α	0.199	0.247	0.553	0.118	0.256	0.330
β	0.580	0.608	0.005*	0.478	0.106	0.498
MAE(1)	1.11	1.03	0.87	0.97	0.91	0.66
MAE(20)	1.06	1.08	0.91	1.06	2.80	0.92
MAE(60)	1.10	1.08	0.93	1.08	4.24	0.96

Note: The sample consists of the daily S&P 500 index (sp), Dow Jones Industrial Average (dj), CRSP equally weighted index (cre), CRSP value-weighted index (crv), federal funds rate (ffr), and yen per dollar exchange rate (yd) from January 4, 1988 through October 6, 2003. The MAEs for the holding sample October 7, 2003 through December 31, 2003 are expressed in percent of the GARCH(1, 1) error as reported in Table 1.

*Indicates the only parameter estimate that is not significant according to all common confidence levels.

The interpretation of persistence in a FIGARCH model is difficult. The estimates of the parameter of fractional integration are highly significant across all series. Davidson (2004) note, however, that in the FIGARCH model the parameter δ does not have the same interpretation as d in the ARFIMA model. The FIGARCH process is non-stationary if $\delta>0$, thus our estimates indicate non-stationarity rather than stationary long memory. Also, the autoregressive parameters α and β do not have a straightforward interpretation as persistence parameters. Their sums are substantially below one for all considered series but we cannot extract a persistence time scale from this as in the case of GARCH.

In summary, long-memory models describe financial volatility data well and the parameter of fractional integration is usually highly significant. This may indicate that there are several persistence time scales present in the data. Fractionally integrated models do not allow, however, for an identification of the different scales.

4.2. Two-Scale GARCH

In the GARCH framework, Engle and Lee (1999) proposed a model that allows for two different overlaying persistence structures in volatility.According to (3), the conditional variance in the GARCH(1, 1) model is given by

$$h_t=\sigma^2+\alpha(\varepsilon_{t-1}^2-\sigma^2)+\beta(h_{t-1}-\sigma^2).$$

Engle and Lee (1999) generalize the unconditional variance σ^2 from a constant to a time varying process q_t, which is supposed to model the highly persistent component of volatility:

$$h_t - q_t = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \beta(h_{t-1} - q_{t-1}), \tag{14}$$

$$q_t = \omega + \rho q_{t-1} + \phi(\varepsilon_{t-1}^2 - h_{t-1}). \tag{15}$$

Then, $\lambda = \alpha + \beta$ can capture the short time scale. The authors show that the model can be written as a GARCH(2, 2) process.

Table 9 reports the estimation of the Engle and Lee (1999) model on the time series considered in this paper. The results clearly show two different scales. The sum $\alpha + \beta$ of the autoregressive parameters indicates a short scale between 5 and 18 days. The autoregressive coefficient ρ of the slowly moving component indicates very high persistence close to integration. As before, the last three rows report the mean absolute forecast errors for the holding period as a fraction of the errors of the GARCH(1, 1) forecast. Except for two instances at the 1-day horizon, the forecast errors are higher than those of the GARCH(1, 1) model.

Table 9. Component-GARCH Estimation.

	sp	dj	cre	crv	ffr	yd
$h_t - q_t = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \beta(h_{t-1} - q_{t-1})$, $q_t = \omega + pq_{t-1} + \phi(\varepsilon_{t-1}^2 - h_{t-1})$.						
α	0.056***	0.061***	0.199***	0.074***	0.139**	0.050
	(0.013)	(0.016)	(0.025)	(0.006)	(0.048)	(0.024)
β	0.888***	0.882***	0.717***	0.861***	0.700***	0.726***
	(0.027)	(0.030)	(0.031)	(0.030)	(0.031)	(0.140)
ρ	0.999***	0.999***	0.999***	0.999***	0.999***	0.986***
	(2e-4)	(2e-4)	(3e-4)	(3e-4)	(6e-4)	(0.008)
ϕ	0.011	0.009	0.014	0.011	0.088	0.028
	(0.255)	(0.313)	(0.181)	(0.221)	(0.189)	(0.026)
ω	2e-5***	2e-5***	2e-5***	2e-5***	0.001**	2e-4**
	(6e-6)	(6e-6)	(5e-6)	(6e-6)	(6e-4)	(1e-4)
MAE(1)	1.58	1.80	1.09	1.37	0.87	0.87
MAE(20)	2.02	2.25	1.34	1.87	2.35	1.08
MAE(60)	2.78	2.95	1.42	2.40	4.26	1.24

Note: Estimation of the Engle and Lee (1999) model for the daily S&P 500 index (sp), Dow Jones Industrial Average (dj), CRSP equally weighted index (cre), CRSP value-weighted index (crv), federal funds rate (ffr), and yen per dollar exchange rate (yd) from January 4, 1988 through October 6, 2003. The MAEs for the holding sample October 7, 2003 through December 31, 2003 are expressed in percent of the GARCH(1, 1) error as reported in Table 1.

**Significant at 95% level;

***Significant at 99% level.

4.3. Wavelet Analysis and Heterogeneous ARCH

The presence of different time scales in a time series (or "signal") is a well-studied subject in the physical sciences. A standard approach to the problem is spectral analysis, that is, Fourier decomposition of the time series. The periodogram estimation carried out in Section 2.3 rests on such a Fourier decomposition. The time series $x(t)$ is assumed to be a stationary process and for ease of presentation, assumed to be defined on continuous time. Then, the process is decomposed into a sum of sines and cosines with different frequencies ω and amplitudes $\hat{x}(\omega)$:

$$x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{x}(\omega) e^{i\omega t} \mathrm{d}\omega. \tag{16}$$

The amplitudes $\hat{x}(\omega)$ are given by the Fourier transform of $x(t)$:

$$x(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(\omega) e^{-i\omega t} \mathrm{d}t. \tag{17}$$

The periodograms in Figs. 1 through 3 are essentially plots of the $\hat{x}(\omega)$ against the frequencies ω.

We saw that spectral theory can be used to identify the short scale. However, this result pertains to the entire sample period only. The integration over the time dimension in (17) compounds all local information into a global number, the amplitude $\hat{x}(\omega)$. The sinusoids $e^{i\omega t} = cos\,\omega t + i\,sin\,\omega t$ have support on all $t \in \mathbb{R}$. The coefficient in the Fourier decomposition (16), the amplitude $\hat{x}(\omega)$, has therefore only global meaning.

It is an interesting question, however, if there are time scales in the data that are of local importance, that is, influence the process for some time and then die out or are replaced by other scales. This problem can be approached using the wavelet transform instead of the Fourier transform (Mallat, 1999). The wavelet transform decomposes a process into wavelet functions which have a frequency parameter, now called scale, and also a position parameter. Thereby, the wavelet transform results in coefficients which indicate a time scale *plus* a position in time where this time scale was relevant. Contrary to that, the Fourier coefficients indicate a globally relevant time scale (frequency) only.

A wavelet $\psi(t)$ is a function with mean zero:

$$\int_{-\infty}^{\infty} \psi(t) \mathrm{d}t = 0,$$

which is dilated with a scale parameter s (corresponding to the frequency in Fourier analysis), and translated by u (this is a new parameter that captures the point in time where the scale is relevant):

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right).$$

In analogy to the Fourier decomposition (16), the wavelet decomposition of the process $x(t)$ is given by

$$x(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Wx(u,s)\psi_{u,s}(t)\mathrm{d}u\,\mathrm{d}s. \tag{18}$$

The coefficients $Wx(u, s)$ are given by the wavelet transform of $x(t)$ at the scale s and position u, which is calculated by convoluting $x(t)$ with the complex conjugate of the wavelet function:

$$Wx(u,s) = \int_{-\infty}^{\infty} x(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)\mathrm{d}t. \tag{19}$$

This is the analog to the Fourier transform (17).

Figures 4 and 5 plot the wavelet coefficients $Wx(u, s)$ against position $u \in \{1,\ldots, 4037\}$ in the time series and against time scale $s \in \{2, 4,..., 512\}$ days. The wavelet function used here is of the Daubechies class with 4 vanishing moments (Mallat, 1999, 249 ff.) but other wavelet functions yield very similar results. The plots for the CRSP data are left out for brevity, they are very similar to Fig. 4. These are three dimensional graphs; the figures present a bird's eye view of the coefficient surface. The color encodes the size of the coefficient *Wx(u, s):* the higher the coefficient, the brighter its color.[3]

If there were a long scale and a short scale in the data, which influenced the process continuously and simultaneously, we would expect to see two horizontal ridges, one at the long and one at the short scale. According to the GARCH estimation in Table 1, the ridge of the long scale of the S&P 500 should be located at $s = 1/(1–0.995) = 200$ days, in the case of the Dow Jones at $s = 1/(1–0.99) = 100$ days, and so on. From the results in Section 2.3, we would expect the ridge of the short scale to be located at around 5 days.

Figures 4 and 5 do not exhibit such horizontal ridges. Instead, there are vertical ridges and braid-like bulges that run from longer scales to shorter scales. A pronounced example is in the upper panel of Fig. 5, where a diagonal bulge runs from observation 500 and the long end of the scale axis to observation 1250 and the short end of the scale axis. A high vertical ridge is located around observation 750. The corresponding event in the federal

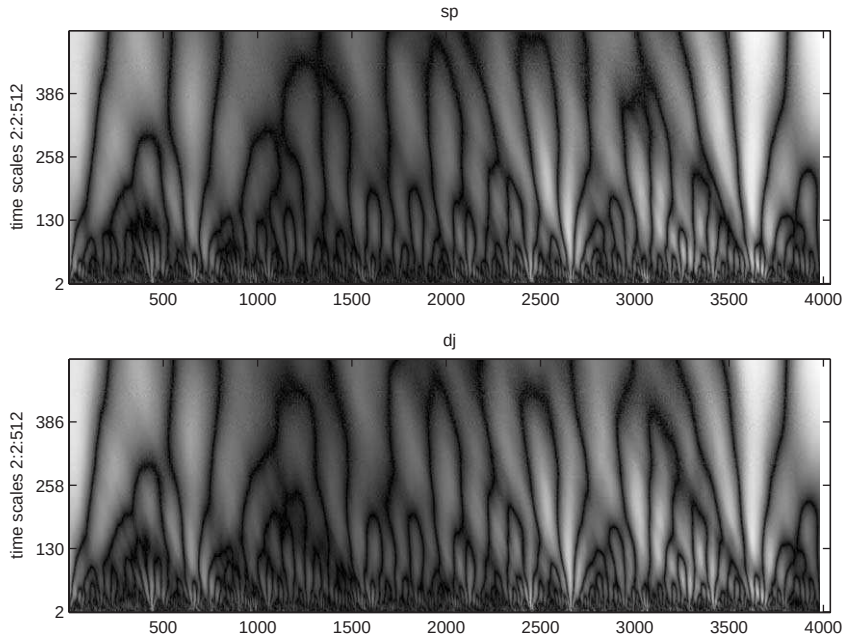


Fig. 4. Wavelet Estimation I. Plot of the Wavelet Coefficients of the Absolute Returns Series of the S&P 500 and the Dow Jones. The Wavelet Coefficients $Wx(u, s)$ are Defined for a Specific Position u, here on the Abscissa, and a Specific Time Scale s, Here on the Ordinate. The Panels give a Bird's Eye View of a Three Dimensional Graph that Plots the Coefficients Against Position and Scale. The Different Values of the Coefficients are Indicated by the Color: the Brighter the Color, the Higher the Coefficient at that Position and Scale.

funds rate series is the period of high fluctuations during the end of the year 1990 and early 1991, when the Federal Reserve phased out of minimum reserve requirements on non-transaction funds and the first gulf war was impending.

These patterns indicate that there is correlation between fluctuations with long-mean reversion time and fluctuations with short-mean reversion time. This correlation has been established by Müller et al. (1997). Ghashgaie et al. (1996) relate this finding to hydrodynamic turbulence, where energy flows from long scales to short scales. Müller et al. (1997) propose a variant of ARCH, called HARCH, to account for this correlation. The idea is to use long-term fluctuations to forecast short-term volatility. This is achieved by adding squared j-day returns to the conditional variance equation. That is,

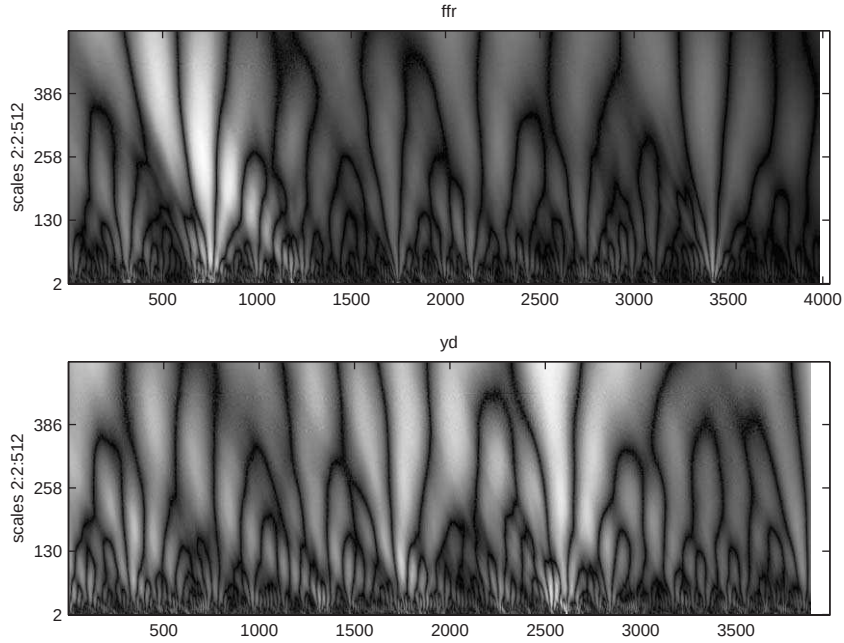


Fig. 5. Wavelet Estimation II. Plot of the Wavelet Coefficients of the Absolute Returns Series of the Federal Funds Rate and the Yen–Dollar Exchange Rate. The Wavelet Coefficients $Wx(u, s)$ are Defined for a Specific Position u, here on the Abscissa, and a Specific Time Scale s, here on the Ordinate. The Panels give a Bird's Eye View of a Three Dimensional Graph that Plots the Coefficients Against Position and Scale. The Different Values of the Coefficients are Indicated by the Color: the Brighter the Color, the Higher the Coefficient at that Position and Scale.

(2) is replaced by

$$h_t = \omega + \sum_{j=1}^{n} \alpha_j \left(\sum_{i=1}^{j} \varepsilon_{t-i} \right)^2, \tag{20}$$

where all coefficients are positive. The number of possible combinations of n and j renders information criterion searches for the best specification infeasible, in particular, since high js are desirable to capture the long-mean reversion scales. The stationarity condition is $\lambda = \Sigma_{j=1}^{n} j\alpha_j < 1$. Müller et al. (1997) also propose a GARCH-like generalization of HARCH, adding lagged values of h to (20). We refrain from using such a specification, since it incurs spurious persistence estimates due to change-points (Hillebrand, 2004).

Table 10. HARCH Estimation.

	sp	dj	cre	crv	ffr	yd
$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2(\Sigma_{i=1}^{20}\varepsilon_{t-i})^2 + \alpha_3(\Sigma_{i=1}^{60}\varepsilon_{t-i})^2 + \alpha_4(\Sigma_{i=1}^{100}\varepsilon_{t-i})^2 + \alpha_5(\Sigma_{i=1}^{250}\varepsilon_{t-i})^2$						
ω	6e-5***	6e-5***	2e-5***	5e-5***	0.047***	3e-5***
	(2e-7)	(2e-46)	(7e-7)	(2e-6)	(3e-4)	(8e-7)
α_1	0.132***	0.120***	0.327***	0.153***	0.556***	0.089***
	(0.013)	(0.012)	(0.023)	(0.009)	(0.011)	(0.009)
α_2	0.010***	0.012***	0.003***	0.017***	1e-10	0.007***
	(0.001)	(9e-4)	(3e-4)	(0.011)	(4e-4)	(8e-4)
α_3	0.002***	0.001***	4e-4***	0.002***	1e-10	0.001***
	(3e-4)	(3e-4)	(6e-5)	(3e-4)	(4e-4)	(2e-4)
α_4	1.5e-4	1e-10	1e-10	3e-4***	1e-10	2e-4***
	(9.5e-5)	(1e-4)	(2e-5)	(8e-5)	(1e-4)	(9e-5)
α_5	4.5e-6	1e-5***	1e-10	1e-10	1e-5***	1e-10
	(3.9e-6)	(4e-6)	(3e-6)	(4e-6)	(5e-6)	(5e-6)
λ	0.45	0.43	0.42	0.49	0.56	0.30
MAE(1)	0.62	0.62	0.25	0.50	1.84	1.59
MAE(20)	1.00	1.16	0.88	0.99	2.51	1.31
MAE(60)	1.17	1.31	0.92	1.09	2.83	1.23

Note: The sample consists of the daily S&P 500 index (sp), Dow Jones Industrial Average (dj), CRSP equally-weighted index (cre), CRSP value-weighted index (crv), federal funds rate (ffr), and yen per dollar exchange rate (yd) from January 4, 1988 through October 6, 2003. The MAEs for the holding sample October 7, 2003 through December 31, 2003 are expressed in percent of the GARCH(1, 1) error as reported in Table 1.
***Significant at 99% level.

We use $n = 5$, that is, we include five different returns at horizons 1, 20, 60, 100, and 250 days. Thereby, we nest the simple ARCH(1) specification but also let 1-month, 3-month, 5-month, and 1-year returns influence daily fluctuations. Table 10 reports the estimation results. Remarkable are the consistently high significance of 3-month returns and the low estimates of the persistence parameter λ. HARCH yields consistent improvements on the 1-day forecast horizon for the four stock-market series, but performs badly on the federal funds rate and the exchange rate.

5. CONCLUSION

This study shows that apart from the well-known high persistence in financial volatility, there is also a short correlation, or fast mean reverting time scale. We find it in six different daily financial time series: four capture the U.S. stock market, the federal funds rate and the yen–dollar exchange rate.

This short scale has a mean reversion time of less than 20 days. An ARMA fit to the residual $\varepsilon_t^2 - h_t$ from a GARCH estimation improves the short-term forecast of the GARCH model and provides a benchmark.

We estimate several generalizations of GARCH that allow for multiple correlation time scales, including segmentation of the data and local GARCH estimation. Unfortunately, none of the considered models is able to fully exploit the information contained in the short scale and improve over the benchmark from the simple ARMA fit.

Wavelet analysis of the volatility time series reveals that there is correlation between fluctuations on long scales and fluctuations on short scales. The heterogeneous ARCH model of Müller et al. (1997), which allows for correlation of this kind, can exploit some of the short-scale information from the stock market part of our data set.

NOTES

1. The relation between the mean reversion time $1/c$ estimated from the Lorentzian and the mean reversion time $1/(1-\lambda)$ from the GARCH model was found in simulations as $1/c = -86.74 + 61.20 \cdot 1/(1-\lambda)$, $R^2 = 0.93$. A motivation of the Lorentzian for GARCH models and the details of the simulations are available upon request.

2. Let $Y_t = \int_{-\infty}^{t} (t-s) H^{-1/2} dW(s)$ be fractional Brownian motion, where $W(t)$ is standard Brownian motion. Then, the relation between the Hurst coefficient H and the fractional integration parameter d is given by $H = d + 1/2$, see Geweke and Porter-Hudak (1983).

3. The color plots can be downloaded at http://www.bus.lsu.edu/economics/faculty/ehillebrand/personal/.

ACKNOWLEDGMENTS

This paper benefited from discussions with and comments from George Papanicolaou, Knut Solna, Mordecai Kurz, Caio Almeida, Doron Levy, Carter Hill, and the participants of the 3rd Annual Advances in Econometrics Conference, in particular Robert Engle and Tom Fomby. Remaining errors of any kind are mine.

REFERENCES

Andersen, T. G., & Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, *4*, 115–158.

Andreou, E., & Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, *17*, 579–600.

Baillie, R. T., Bollerslev, T., & 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.

Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modeling in finance: A review of theory and empirical evidence. *Journal of Econometrics*, *52*, 5–59.

Bollerslev, T., & Engle, R. F. (1993). Common persistence in conditional variances. *Econometrica*, *61*(1), 167–186.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). GARCH models. In: R. F. Engle & D. L. McFadden (Eds), *Handbook of Econometrics* (Vol. 4). Amsterdam: Elsevier.

Chernov, M., Gallant, A. R., Ghysels, E., & Tauchen, G. (2003). Alternative models of stock-price dynamics. *Journal of Econometrics*, *116*, 225–257.

Davidson, J. (2004). Moment and memory properties of linear conditional heteroskedasticity models, and a new model. *Journal of Business and Economics Statistics*, *22*(1), 16–29.

Den Haan, W., & Levin, A. (1997). A practitioner's guide to robust covariance matrix estimation. In: G. Maddala & C. Rao, (Eds), *Handbook of Statistics* (Vol. 15). Amsterdam, North-Holland: Robust Inference.

Diebold, F. X. (1986). Modeling the persistence of conditional variances: A comment. *Econometric Reviews*, *5*, 51–56.

Ding, Z., & 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., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*, 83–106.

Doornik, J. A. (2002). *Object-oriented matrix programming using Ox* (3rd ed.). London: Timberlake Consultants Press. www.nuff.ox.ac.uk/Users/Doornik

Doornik, J. A., & Ooms, M. (1999). A package for estimating, forecasting and simulating arfima models. www.nuff.ox.ac.uk/Users/Doornik

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007.

Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, *5*(1), 1–50.

Engle, R. F., & Lee, G. G. J. (1999). A long-run and short-run component model of stock return volatility. In: R. F. Engle & H. White (Eds), *Cointegration, causality, and forecasting: A festschrift in honour of Clive W. J. Granger*. Oxford: Oxford University Press.

Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance*, *1*(2), 237–245.

Fouque, J. P., Papanicolaou, G., Sircar, K. R., & Sølna, K. (2003). Short time-scale in S&P 500 volatility. *Journal of Computational Finance*, *6*, 1–23.

Gallant, A. R., & Tauchen, G. (2001). *Efficient method of moments*. Mimeo. http://www.unc.edu/∼arg

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, *4*(4), 221–238.

Ghashgaie, S., Breymann, W., Peinke, J., Talkner, P., & Dodge, Y. (1996). Turbulent cascades in foreign exchange markets. *Nature*, *381*, 767–770.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, *14*, 227–238.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, *1*(1), 15–29.

Hamilton, J. D., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*, 307–333.

Hillebrand, E. (2004). *Neglecting parameter changes in autoregressive models*. Louisiana State University Working Paper.

Hillebrand, E. (2005). Neglecting Parameter Changes in GARCH Models. *Journal of Econometrics,* 129, Annals Issue on Modeling Structural Breaks, Long Memory, and Stock Market Volatility (forthcoming).

Kokoszka, P., & Leipus, R. (1999). Testing for parameter changes in ARCH models. *Lithuanian Mathematical Journal*, *39*, 182–195.

Lamoureux, C. G., & Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business and Economic Statistics*, *8*, 225–234.

Laurent, S., & Peters, J.-P. (2002). Garch 2.2: An Ox package for estimating and forecasting various ARCH models. *Journal of Economic Surveys*, *16*(3), 447–485.

LeBaron, B. (2001). Stochastic volatility as a simple generator of apparent financial power laws and long memory. *Quantitative Finance*, *1*(6), 621–631.

Mallat, S. (1999). *A wavelet tour of signal processing* (2nd ed.). San Diego: Academic Press.

Mikosch, T., & Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics*, *86*(1), 378–390.

Müller, U. A., Dacorogna, M. M., Dave, R. D., Olsen, R. B., Pictet, O. V., & von Weizsaecker, J. E. (1997). Volatilities of different time resolutions – analyzing the dynamics of market components. *Journal of Empirical Finance*, *4*, 213–239.

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

EVALUATING THE 'FED MODEL' OF STOCK PRICE VALUATION: AN OUT-OF-SAMPLE FORECASTING PERSPECTIVE

Dennis W. Jansen and Zijun Wang

ABSTRACT

The "Fed Model" postulates a cointegrating relationship between the equity yield on the S&P 500 and the bond yield. We evaluate the Fed Model as a vector error correction forecasting model for stock prices and for bond yields. We compare out-of-sample forecasts of each of these two variables from a univariate model and various versions of the Fed Model including both linear and nonlinear vector error correction models. We find that for stock prices the Fed Model improves on the univariate model for longer-horizon forecasts, and the nonlinear vector error correction model performs even better than its linear version.

1. INTRODUCTION

A simple model, the so-called "Fed Model," has been proposed for predicting stock returns or the level of the stock market as measured by the S&P 500 index. This model compares the earnings yield on the S&P 500 to

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 179–204
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20026-9

the yield on 10-year government bonds, where the earnings yield on the S&P 500 is calculated as the predicted future earnings of S&P 500 firms divided by the current value of the S&P 500 index. This model is not endorsed by the Board of Governors of the Federal Reserve System (i.e. the Fed), but it was used in the Fed's Humphrey–Hawkins report to Congress in July 1997 and quickly picked up by private security analysts, who gave it the name "Fed Model." Variants have been adopted by a number of security strategists including some at for example Prudential Securities and Morgan Stanley.

The basic idea behind the Fed Model is that the yields on the S&P 500 and on the 10-year government bond should be approximately the same, and certainly should move together over time, so that deviations of the equity yield from the bond yield signal that one or both yields must adjust to restore the long-run equilibrium relationship between the two yields.

A simple textbook asset pricing formula would link the value of a perpetuity to the ratio of the annual payment to the interest rate, after which a simple manipulation would lead to the equality of the interest rate and the asset yield, the equilibrium relationship postulated by the Fed Model. The Fed Model can be seen as postulating this simple asset pricing formula as a long-run equilibrium relationship instead of a relationship that holds observation by observation.

The Fed Model is not, however, completely consistent with the simple textbook asset pricing formula. The textbook asset pricing formula would have equality between the real interest rate and the asset yield, not equality between the nominal interest rate and the asset yield. Nonetheless, applications of the Fed Model have ignored the distinction between real and nominal interest rates (see Asness, 2003).

In this paper, we propose to investigate the Fed Model, to consider its statistical properties including stability over time, and especially to consider the predictive power of the Fed Model. We ask if the Fed Model contains useful information for forecasting stock prices or bond yields. We conduct an out-of-sample forecasting exercise and investigate the forecasting performance of the Fed Model at various horizons.

Since the Fed Model can be interpreted as postulating a cointegrating relationship between the equity yield and the bond yield, our investigation will provide insights into the usefulness of incorporating this cointegrating relationship in a forecast of equity and bond yields. Further, because the postulated cointegrating relationship is a long-run equilibrium relationship, we evaluate the Fed Model's forecasting ability for various forecasting horizons and thus provide information on the relative utility of incorporating the cointegrating relationship into short- and longer-run forecasts.

We further investigate the predictive ability of the Fed Model in a nonlinear framework. Interpreting the Fed Model as postulating the cointegrating relationship between the equity and bond yield, we ask if a nonlinear threshold cointegration approach can improve predictive performance over and above the linear version of the Fed Model. Kilian and Taylor (2003) argue that nonlinear mean reversion better describes asset price movements in a world of noise trading and risky arbitrage. They hypothesize that the risk to arbitrage decreases as assets become increasingly over or under priced, leading to quicker adjustments toward equilibrium. Thus small differences are likely to be persistent, while large deviations are more quickly corrected. Kilian and Taylor (2003) suggest an exponential smooth transition autoregressive (ESTAR) model for nominal exchange rate deviations from purchasing power parity, while Rapach and Wohar (2005) consider a similar ESTAR model of the price-dividend and price-earnings ratios.

2. DATA

We obtained monthly data from January 1979 through February 2004 on the average of analysts' estimated 12-month forward earnings on the S&P 500 (*FE*), the S&P 500 price index (*SP*) and the 10-year U.S. Treasury bond yield (*BY*), all transformed to natural logarithms. This study was limited to a January 1979 starting point because this was the earliest period for which we obtained estimated forward earnings. These estimated forward earnings are estimates from private analysts and are not available over the much longer time span for which data on the S&P 500 is available.

For purposes of conducting an out-of-sample forecasting exercise we divided this sample into two parts, a subsample to be used for estimation, and a subsample to be used for out-of-sample forecast evaluation. The subsample for model specification and estimation is January 1979–February 1994, and the forecast evaluation period is March 1994–February 2004. We chose the forecast evaluation period to be a 10-year period because we wanted a relatively long period in order to evaluate longer-horizon forecasts and in order to avoid having the late-1990s stock market boom and subsequent bust to unduly influence our results.

Panel A of Fig. 1 plots the three series *FE*, *SP* and *BY* for our entire sample period. All three series are measured in natural logarithms, and *FE* was multiplied by 100 (prior to taking logs) for ease of presentation. Both *SP* and *FE* show a general upward trend for most of the period, until

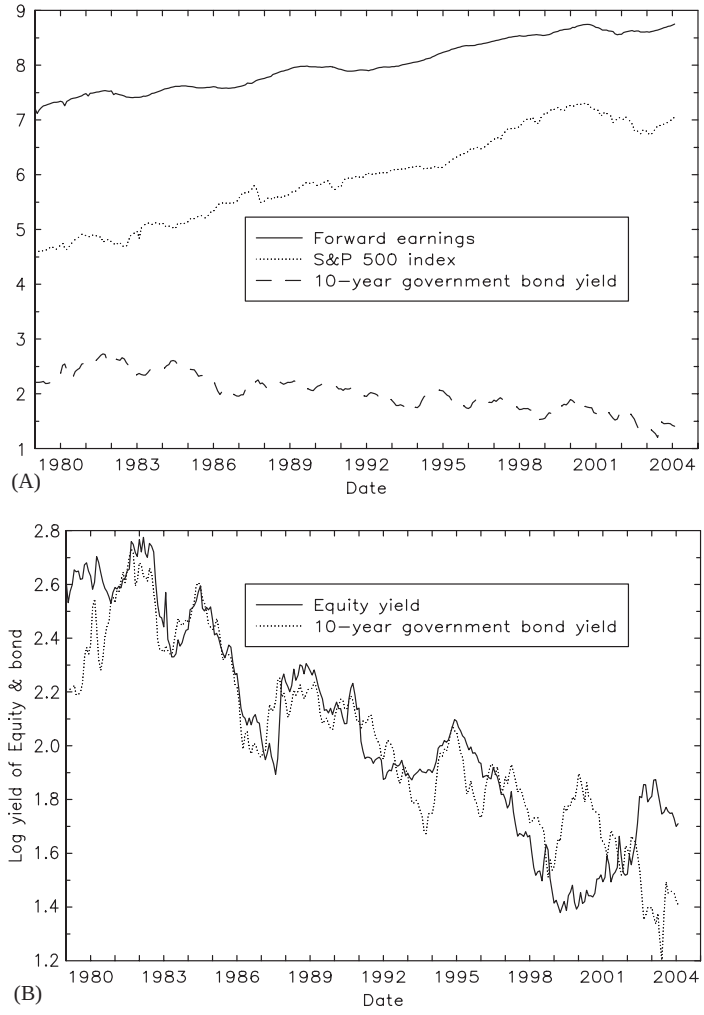


Fig. 1. Forward Earnings, S&P 500 Index and 10-Year Government Bond Yield. Panel (A) and Panel (B).

September 2000 for *FE* and August 2000 for *SP*. There is a less significant downward trend in *BY* over the sample period.

We defined the equity yield (*EY*) as the difference between *FE* and *SP*. Panel B of Fig. 1 plots the equity yield, *EY*, and the bond yield, *BY*. Over the entire sample period of January 1979–February 2004, the average stock

yield is 2.06 (log percentage points), slightly higher than the average bond yield of 2.03. From Fig. 1 it appears that the two yield series move together over the sample period, but there is a greater tendency for short-run divergence in the very beginning of the sample, and again beginning in 1999.

Table 1 summarizes the results of unit root tests on each series. The first half of Table 1 presents the results over the initial estimation and specification subsample, January 1979–February 1994. Over this period we do not reject the null of a unit root in any of the four series, FE, SP, BY or EY, regardless of our trend specification. We do, however, reject a unit root in the yield difference, $EY-BY$, providing support for the Fed Model's implication that these variables are cointegrated, and with a specific cointegrating vector.

For completeness, the second half of Table 1 presents unit root results over the entire sample period of January 1979–February 2004. Here the

Table 1. Results of the Augmented-Dickey–Fuller (ADF) Test.

Variable Name	Including Drift		Including Drift and Trend	
	Lag order	*t*-statistic	Lag order	*t*-statistic
Estimation subsample: 1979.01–1994.02				
Forward earnings	0	−0.542	4	−2.579
Stock price	0	−0.694	1	−3.323
Bond yield	2	−0.501	2	−2.961
Equity yield	0	−1.079	1	−2.736
Yield difference	1	−4.635**	—	—
Full sample: 1979.01–2004.02				
Forward earnings	3	−0.370	3	−2.510
Stock price	1	−1.032	1	−1.833
Bond yield	3	−0.895	3	−4.212**
Equity yield	1	−1.521	1	−2.340
Yield difference	1	−3.403*	—	—

Notes:(1) Forward earnings (FE), stock price (SP) and bond yield (BY) are all in logs. Equity yield is defined as $FE - SP$. The yield difference is the difference of the equity yield and the bond yield, $FE - SP - BY$. (2) The lag order for the tests was determined by the Schwarz information criterion (SIC) in a search over a maximum of 12 lags. (3) MacKinnon critical values for 170 sample observations (corresponding to our estimation subsample) are: test with drift, −3.470 (1%) and −2.878 (5%); test with drift and trend −4.014 (1%) and −3.437 (5%). For 290 sample observations (corresponding to our full sample) critical values are: test with drift, −3.455 (1%) and −2.872 (5%); test with drift and trend, −3.993 (1%) and −3.427 (5%).

*Indicates rejection of null at 5% critical value;

**indicates rejection of null at 1% critical value.

results are much the same as over the subsample, except that when the alternative hypothesis allows a drift we would reject the unit root null for the bond yield variable, BY. For purposes of considering the Fed Model, we note that for the yield difference $EY-BY$ we again reject a unit root. Thus the preliminary indication is that the main hypothesis drawn from the Fed Model, that the yield difference is stationary, holds over the estimation subsample and over the full sample.

In first differences, all four of our series reject the null of a unit root over either the estimation subsample or the full sample.

3. ECONOMETRIC METHODOLOGY

3.1. Linear Error Correction Models

The basic empirical framework used in this study is the vector error correction, or VECM, model. Let Y_t denote a (3×1) vector, $Y_t = (y_{1,t}, y_{2,t}, y_{3,t})'$. Here, y_1 stands for the forward earnings (FE), y_2 for stock price index (SP) and y_3 for the bond yield (BY). Using conventional notation, this model can be described as

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \mu + \varepsilon_t, \quad t = 1, \ldots, T \tag{1}$$

and the error correction form of Eq. (1) is

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + \mu + \varepsilon_t, \quad t = 1, ..., T \tag{2}$$

with

$$\Gamma_i = -(A_{i+1} + A_{i+2} + \cdots + A_p), \quad i = 1, 2, \ldots, p-1$$

and

$$\Pi = -(I_m - A_1 - A_2 - \cdots - A_p)$$

where Δ is the difference operator ($\Delta Y_t = Y_t - Y_{t-1}$), Π a (3×3) coefficient matrix, Γ_i $(i = 1, 2, \ldots, p-1)$ a (3×3) matrix of short-run dynamics coefficients, μ a (3×1) vector of parameters and ε a vector of Gaussian errors. If the rank of matrix Π, r, is 0, then the three variables are not cointegrated. A VAR in first differences of the three variables, ΔFE, ΔSP and ΔBY, is appropriate. If $0 < r < 3$, then the three variables are cointegrated, implying that there exists a long-run relationship between the

three series. In this case, Π can be written as $\Pi = \alpha\beta'$, where α and β' are of the dimension $3 \times r$. Then β is called the cointegrating vector, and reflects the long-run relationship between the series. The term $\beta' Y_{t-1}$ is the error correction term, while α a measure of the average speed by which the variables converge toward the long-run cointegrating relationship. In particular, if the stock yield equals the bond yield in the long run, as the Fed model suggests, then a natural candidate for the cointegrating vector in Eq. (2) is $FE-SP-BY$.

To examine whether the inclusion of series y_j in the regression of y_i ($j \neq i$) helps improve predictability of series y_i, we compare the out-of-sample forecasting performance of the trivariate model given by Eq. (2) (hereafter the VECM model) with that of the following univariate model:

$$\Delta y_{i,t} = \sum_{j=1}^{p-1} \gamma_{ij} \Delta y_{i,t-j} + \mu_i + \varepsilon_{i,t}, \quad t = 1, \ldots, T, \quad i = 1, 2 \tag{3}$$

Note that in this model, the lag terms of Δy_j are not explanatory variables for Δy_i ($i \neq j$), nor is the error correction term $\beta' Y_{t-1}$.

3.2. Linearity Testing and STAR Model Specification

The above linear error correction model in Eq. (2) has several limitations. Two that we will examine are, first, that it does not allow the possibility that error correction or dynamic adjustment might only occur if the difference in yields on the two securities exceeds a certain threshold (perhaps due to transaction costs). Second, the speed of adjustment toward the long-run cointegrating relationship might vary with the magnitude of the difference in yields on the two securities, perhaps due to heterogeneous behavior of investors. While Eq. (2) does not allow for these possibilities, both can be allowed in a smooth transition autoregressive (STAR) model, assuming SY and BY are cointegrated:

$$\begin{aligned} \Delta y_{i,t} = & \sum_{j=1}^{p-1} \sum_{k=1}^{2} \gamma_{ijk} \Delta y_{k,t-j} + \alpha_i EC_{t-1} + \mu_i \\ & + \left(\sum_{j=1}^{p-1} \sum_{k=1}^{2} \gamma^*_{ijk} \Delta y_{k,t-j} + \alpha^*_i EC_{t-1} + \mu^*_i \right) F(z_{t-d}) + \varepsilon_{i,t}, \quad i = 1, 2 \end{aligned} \tag{4}$$

where EC_{t-1} is the error correction term, $EC_{t-1} = \beta' Y_{t-1}$, z_{t-d} the transition variable and d the delay parameter, the lag between a change in the transition

variable and the resulting switch in dynamics.[1] The function $F(\cdot)$ varies from zero to one and determines the switch in dynamics across regimes.

In practice, there are two main specifications of the transition function. One is the logistic function:

$$F(z_{t-d}) = (1 + e^{-\theta(z_{t-d}-c)})^{-1} \tag{5}$$

Models using this transition function are often labeled logistic smooth transition autoregressive or LSTAR models.

The alternative specification is the exponential function:

$$F(z_{t-d}) = 1 - e^{-\theta(z_{t-d}-c)^2} \tag{6}$$

Models using this transition function are often called exponential smooth transition autoregressive or ESTAR models. For both Eq. (5) and (6), the parameter $\theta > 0$ measures the speed of transition between regimes.

In this paper, the transition variable is the error correction term, so $z_{t-d} = \beta' Y_{t-d}$. Therefore, the dynamics of our three-variable system change as a function of the deviation of our three variables from the specified long-run cointegrating relationship.

To conduct the linearity test, we follow the procedure proposed by Terasvirta and Anderson (1992). Specifically, we estimated the following equation:

$$\begin{aligned}\Delta y_{i,t} = {} & \sum_{j=1}^{p-1}\sum_{k=1}^{2} \gamma_{ijk}\Delta y_{k,t-j} + \alpha_i EC_{t-1} + \mu_i \\ & + \left(\sum_{j=1}^{p-1}\sum_{k=1}^{2} \gamma^1_{ijk}\Delta y_{k,t-j} + \alpha^1_i EC_{t-1}\right) z_{t-d} \\ & + \left(\sum_{j=1}^{p-1}\sum_{k=1}^{2} \gamma^2_{ijk}\Delta y_{k,t-j} + \alpha^2_i EC_{t-1}\right) z^2_{t-d} \\ & + \left(\sum_{j=1}^{p-1}\sum_{k=1}^{2} \gamma^3_{ijk}\Delta y_{k,t-j} + \alpha^3_i EC_{t-1}\right) z^3_{t-d} + v_{i,t}, \quad i = 1, 2\end{aligned} \tag{7}$$

The value of d is determined through the estimation of (7) for a variety of d values. For each d, one can test the hypothesis that $\gamma^s_{ijk} = \alpha^s_i = 0$ for all j and k, where $s = 1, 2, 3$. If only one value of d indicates rejection of linearity, it becomes the chosen value. On the other hand, if more than one of the d values rejects linearity, Terasvirta (1994) suggests picking the delay that yields the lowest marginal probability value.

Terasvirta and Anderson (1992) also suggest testing the following set of hypothesis:

$$\begin{aligned} H_{0,1} : \gamma_{ijk}^3 &= \alpha_i^3 = 0 \\ H_{0,2} : \gamma_{ijk}^2 &= \alpha_i^2 = 0 | \gamma_{ijk}^3 = \alpha_i^3 = 0 \\ H_{0,3} : \gamma_{ijk}^1 &= \alpha_i^1 = 0 | \gamma_{ijk}^2 = \alpha_i^2 = \gamma_{ijk}^3 = \alpha_i^3 = 0 \end{aligned}$$

Selection of ESTAR and LSTAR can be made using the following decision rules:

(1) If $H_{0,1}$ is rejected, select an LSTAR model.
(2) If $H_{0,1}$ is not rejected impose it. Then if $H_{0,2}$ is rejected, select an ESTAR model.
(3) If neither $H_{0,1}$ nor $H_{0,2}$ are rejected then impose these restrictions. Then if $H_{0,3}$ is rejected, select an LSTAR model.

3.3. Forecast Comparisons

As mentioned above, we divided the data of length T into two subsets, a subsample of length T_1 to be used for model specification and estimation, and a subsample of length T_2 to be used for forecast evaluation. We specified our univariate, linear VECM and nonlinear VECM models based upon the sample information of the first T_1 observations and then generated h-step-ahead recursive forecasts of Y_t for the remaining T_2 observations from all models. We denote the corresponding forecast error series as e_{it}^R and e_{it}^U, respectively. To test whether the forecasts from the various models are statistically different, we apply a test procedure proposed by Diebold and Mariano (1995). The Diebold and Mariano test allows a comparison of the entire distribution of forecast errors from competing models. For a pair of h-step-ahead forecast errors (e_{it}^U *and* e_{it}^R, $t = 1, \ldots, T_2$), the forecast accuracy can be judged based on some specific function $g(\cdot)$ of the forecast error (where mean squared forecast errors, or MSFE, is often used). The null hypothesis of equal forecast performance is:

$$\mathrm{E}[g(e_{it}^U) - g(e_{it}^R)] = 0$$

where $g(\cdot)$ is a loss function. We use $g(u) = u^2$, the square loss function, in this paper.

Define d_t by

$$d_t = g(e_{it}^U) - g(e_{it}^R)$$

Then the Diebold-Mariano test statistic is given by:

$$\mathrm{DM} = [\hat{V}(\bar{d})]^{-1/2}\bar{d}$$

where $\bar{d}$ is the sample mean of d_t and $\hat{V}(\bar{d})$ the Newy–West heteroscedasticity and autoregression consistent estimator of the sample variance of $\bar{d}$. The D–M statistic does not have a standard asymptotic distribution under the null hypothesis when the models being compared are nested (see McCracken (2004) for the one-period-ahead forecasts, and Clark & McCracken (2004) for multiple-period-ahead forecasts). We refer to the simulated critical values provided in Table 1 of McCracken (2004) for the recursive scheme. These critical values apply to the case of one-period-ahead forecasts.

4. EMPIRICAL RESULTS

4.1. Model Specification and Estimation

We use the first 182 monthly observations (January 1979–February 1994) for in-sample model specification, and the remaining 120 monthly observations (10 years) for out-of-sample forecasting comparisons. As the first step in model estimation, we determine the optimum lag order p in Eq. (1). Schwarz information criterion (SIC) is minimized at $p = 2$ while AIC suggests that $p = 4$. Since our interest is out-of-sample forecasts and simple models often deliver better forecasts than complicated models, we choose $p = 2$ in the subsequent analysis.[2]

We test the cointegrating rank of model (2) following Johansen's (1988, 1991) trace procedure (and specifying $p = 2$). The test statistics we calculate are 35.94 and 11.03 for the null hypotheses $r = 0$ and $r \leq 1$, respectively. The asymptotic critical values are 29.80 and 15.49 at the 5% level, so we conclude that $r = 1$. That is, we conclude that there is a single long-run relationship among the three variables, *FE*, *SP* and *BY*. Thus, the first model we consider is a vector error correction model with $p = 2$ (or one lag of the differences).[3]

We conduct two exercises to examine the stability of the long-run cointegrating relationship. First, we conduct a sequence of tests for structural change in which all samples start in March 1979 and the first sample ends in March 1994 the second in April 1994 and so on until the last sample, which ends in February 2004 and includes all 302 observations, our full sample. This exercise is based on the idea that the estimated cointegrating relationship is stable after a possible structural change, and hence including more observations will add to our ability to detect the change.

Second, we conducted a series of tests for structural change in the adjustment coefficients and/or in the cointegrating vector as developed by Seo (1996).

Figure 2 summarizes the results of these exercises. Panel A presents the recursive estimates of cointegration rank, which indicates that the cointegration relationship breaks down temporarily in 1999, is restored in mid-2000, but then breaks down again in mid-2002.[4]

Panels B and C report the sequential break test for the cointegrating vector (Panel B) and the adjustment coefficients (Panel C) as proposed by Seo. Both tests indicate structural change, and there is particularly strong evidence for structural change in the adjustment coefficient.

An advantage of the above vector error form is that it can be easily used to test the interesting hypothesis that the equity yield is equal to the bond yield in the long run, $FE - SP = BY$. This hypothesis can be tested as a restriction on the cointegration space in model (2), a test of $\beta = (1, -1, -1)'$. A simple first check of the appropriateness of this hypothesis is to test the nonstationarity of the yield difference series, $FE - SP - BY$, as we reported in Table 1. This yield difference is stationary over both our estimation subsample and our full sample. When we directly test this restriction within the context of an estimated VECM we find less support, as the Johansen LR test statistic for this hypothesis is 7.81 with a p-value of 0.02.

The restriction of equality between the equity yield and the bond yield, or at least of a stationary and constant difference between the equity yield and the bond yield, is based in part on economic theory. Thus, while the in-sample test result does not support the hypothesis of equal yield, we estimate and forecast a model with this restriction imposed to see whether it can improve the forecasting performance, especially in the long run.[5] This model, an estimated VECM imposing equality between the equity and bond yield, is our second model.

Following the procedure described in Section 3.2, we test for the appropriateness of the linear assumption implicit in the linear models outlined above. For tests of linearity we use the in-sample observations (January 1979 – February 1994). The sequential tests of the null hypotheses $H_{0,1}$, $H_{0,2}$ and $H_{0,3}$ are summarized in Table 2, where we assume the maximum value of d is 3. In conducting these tests, we maintain the restriction of equal yields, so the error correction term is $EC_{t-1} = FE_{t-1} - SP_{t-1} - BY_{t-1}$. The null hypothesis H_0 is that linearity is an appropriate restriction. It can be seen from Table 2 that H_0 can be rejected for all the three equations, that is for FE, SP and BY. Thus, a third forecasting model we consider is a nonlinear VECM. The minimum probability value occurs at

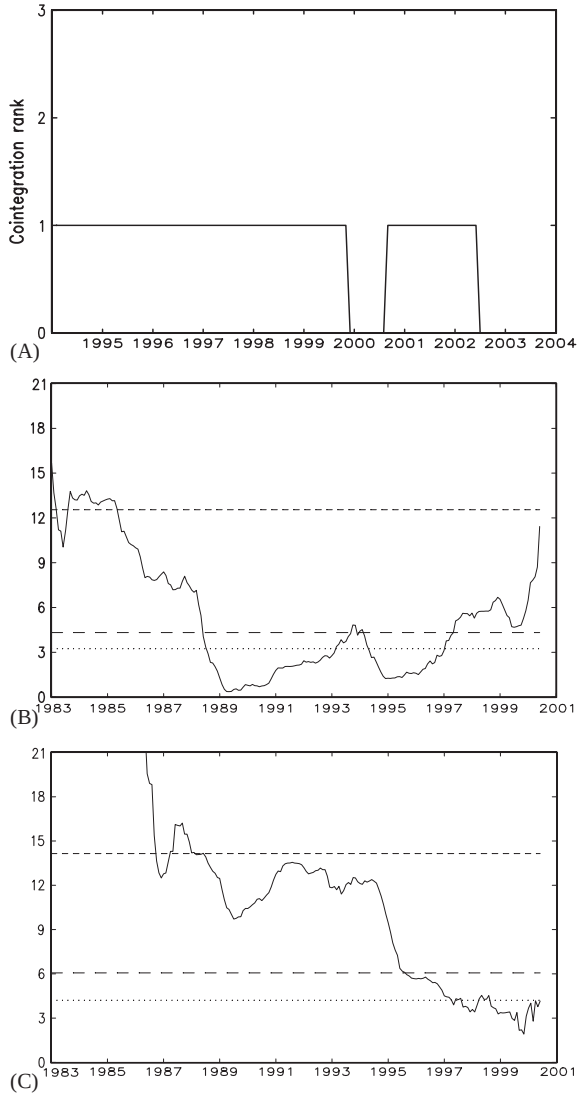


Fig. 2. Stability Tests. Panel (A) Recursive Estimates of Cointegration Rank. Panel (B) Test for Stability of Cointegrating Vector. Panel (C) Test for Stability of Adjustment Coefficient. *Note:* In Panel (A) the First Sample Covers the Period of 1979.03–1994.02, and the Last Sample 1979.03–2004.02. Horizontal Axes Indicate the End Periods for Each Window. The Significance Level is 5%. In Panels B and C, the Short- and Long-Dashed Lines are the 5% Critical Values of the Sup_LM and Ave_LM Test Statistics of Seo (1998).

Table 2. Linearity Tests and Determination of the Delay Parameter d.

Dependent variable	$d = 1$	$d = 2$	$d = 3$
Forward earnings (FE)			
H_0	0.051	0.000	0.000
$H_{0,1}$	0.826	0.015	0.011
$H_{0,2}$	0.454	0.485	0.003
$H_{0,3}$	0.003	0.000	0.001
Stock price (SP)			
H_0	0.000	0.004	0.006
$H_{0,1}$	0.121	0.609	0.278
$H_{0,2}$	0.074	0.016	0.088
$H_{0,3}$	0.001	0.006	0.005
Bond yield (BY)			
H_0	0.008	0.813	0.101
$H_{0,1}$	0.582	0.680	0.027
$H_{0,2}$	0.001	0.420	0.088
$H_{0,3}$	0.351	0.841	0.455

Notes: (1) The four null hypotheses in relation to Eq. (6) are: $H_0 : \gamma_{ijk}^1 = \alpha_i^1 = \gamma_{ijk}^2 = \alpha_i^2 = \gamma_{ijk}^3 = \alpha_i^3 = 0$, $H_{0,1} : \gamma_{ijk}^3 = \alpha_i^3 = 0$, $H_{0,2} : \gamma_{ijk}^2 = \alpha_i^2 = 0 | \gamma_{ijk}^3 = \alpha_i^3 = 0$, $H_{0,3} : \gamma_{ijk}^1 = \alpha_i^1 = 0 | \gamma_{ijk}^2 = \alpha_i^2 = \gamma_{ijk}^3 = \alpha_i^3 = 0$.(2) Column entries are associated p-values.

$d = 3$ for the FE equation and at $d = 1$ for the SP and BY equations. In the SP equation and restricting our attention to $d = 1$, the null hypothesis $H_{0,1}$ is not rejected and $H_{0,2}$ is rejected at the 0.10 but not the 0.05 significance level, while $H_{0,3}$ is rejected even at the 0.01 significance level. Thus, SP could be modeled as either an ESTAR or an LSTAR model with a delay of one. In the BY equation and restricting our attention to $d = 1$, the null hypothesis $H_{0,1}$ is not rejected, but $H_{0,2}$ is rejected, which implies that an ESTAR specification is appropriate for BY. We will use an ESTAR specification for both BY and SP. For FE we have an LSTAR specification with $d = 3$.

To examine whether the above results are sensitive to the choice of the cointegrating rank, we generate forecasts from the linear VECM while imposing the rank restriction $r = 0$. That is, we specify the model as a VAR in first differences. This is our fourth model.

As the cointegration relationship was not stable throughout the forecasting period, we also generate forecasts from models that allow the cointegrating rank to vary across each recursive sample. The fifth model is estimated with increasing-length windows (samples). This allows use to simulate a forecaster in real time estimating the model, testing its

cointegrating rank and making forecasts from the resulting model whether specified as a VECM with $r = 1$ or as a VECM with $r = 0$ (equivalently, a VAR in first differences).

Throughout the paper, we refer to Eq. (3) as the univariate model. It is a simple univariate model in first differences with one or more lags, and will serve as our benchmark model against which the above six multivariate specifications are tested.

We note that a common claim is that stock market prices can be modeled as following a random walk. We also examined the out-of-sample forecasting performance of the random walk model. This model did slightly worse than our univariate model. For *SP* the root mean squared error (RMSE) is 0.0379 for the one-step-ahead forecast and 0.2060 for the 12-step-ahead forecast, both slightly higher than the RMSE values for the univariate model of *SP*. (see Tables 3 and 4 for a comparison.)

Table 3. Out-of-Sample Forecasts (One-Step-Ahead:1994.03–2004.02).

	0	1	2	3	4	5
Tested Model	Linear Univariate AR	Linear VECM; Estimated CV	Linear VECM; Imposed CV	Nonlinear VECM	Linear VAR in Differences	Linear VECM, r varies, Increasing Window
Forecasted variable: stock price (SP)						
RMSE	0.0366	0.0379	0.0378	0.0384	0.0375	0.0378
MAE	0.0281	0.0284	0.0280	0.0285	0.0284	0.0286
D–M statistic	—	−1.467**	−1.146**	−1.701**	−1.461**	—
Forecasted variable: bond yield (BY)						
RMSE	0.0446	0.0445	0.0449	0.0452	0.0447	0.0444
MAE	0.0347	0.0345	0.0349	0.0359	0.0348	0.0346
D–M statistic	—	0.128	−0.256	−0.373*	−0.057	—

Note: Model 0 is a univariate AR model; the comparison null model in each D–M test. Model 1 is a VECM with estimated cointegrating vector. Model 2 is a VECM with imposed cointegrating vector, EC = BY−SY. Model 3 is a nonlinear smooth transition VECM, with ESTAR transition function. Model 4 is a VAR(1) in first differences. Model 5 allows r to vary over increasing-length samples of data.

**significance at the 0.05 level;

*significance at the 0.10 level.

Table 4. Out-of-Sample Forecasts (12-Steps-Ahead: 1995.02–2004.02).

	0	1	2	3	4	5
Tested Model	Linear Univariate AR	Linear VECM; Estimated CV	Linear VECM; Imposed CV	Nonlinear VECM	Linear VAR in Differences	Linear, r varies, Increasing Window
Forecasted variable: stock price (SP)						
RMSE	0.1906	0.1941	0.1826	0.1775	0.1920	0.1994
MAE	0.1540	0.1572	0.1519	0.1468	0.1543	0.1619
D–M statistic	—	−0.400	0.514*	0.679**	−0.627*	
Forecasted variable: bond yield (BY)						
RMSE	0.1659	0.1492	0.1672	0.1606	0.1660	0.1590
MAE	0.1380	0.1122	0.1244	0.1275	0.1388	0.1271
D–M statistic		1.179**	−0.047	0.332*	−0.116	

Note: See Note in Table 3.
**significance at the 0.05 level;
*significance at the 0.10 level.

4.2. Comparisons of Out-of-Sample Forecasts (One-Step-Ahead)

We first recursively estimate the linear VECM model with $p = 2$ and $r = 1$, as well as its restricted version, a univariate model in first differences. From the recursive parameter estimates we generate one-step-ahead up to 12-steps-ahead out-of-sample forecasts for the monthly values of *FE*, *SP* and *BY*. Thus, we generated 120 sets of one-step-ahead forecasts covering the period of March 1994–February 2004, and 109 sets of one-step-ahead forecasts covering the period of February 1995–February 2004. As the forecasts of *FE* are not of direct interest, they are omitted from much of the following discussion. Instead, we focus our attention on our main interest, forecasting stock prices, and our secondary interest, forecasting bond yields.

Forecast errors are derived by subtracting the forecasts from realized values. Two statistics are calculated for each forecast error series, the root mean squared forecast errors (RMSE) and the mean absolute errors (MAE). The results of one-step-ahead forecasts are summarized in Table 3 under the column labeled 1. For *SP*, the RMSE from the VECM is 0.0379, which is larger than the 0.0366 from the univariate model. The MAE from the VECM

is 0.0284, also larger than the univariate model's 0.0281. The D–M statistic is −1.467, indicating that the VECM model generates less accurate forecasts than the univariate model that includes only lags of *SP* at the 0.05 significance level.

The bottom half of Table 3 contains results for forecasts of *BY*. Here, for the column labeled 1 we see that the RMSE and MAE of the VECM model are 0.0445 and 0.0345, respectively, and both are smaller than their counterparts in the univariate model, 0.0446 and 0.0347. Nevertheless, the D–M statistic is insignificant.

In Table 3, in the column labeled 2, we present the forecast comparisons of the VECM with the restricted cointegration space, model 2, and the univariate models. The imposition of this restriction improves the forecasts of *SP* while reducing the forecast accuracy of *BY*. Nevertheless, the forecasts of *SP* from the restricted VECM are still less accurate than the forecasts from the univariate model in terms of RMSE, although they have a slightly smaller MAE than the rival forecasts (0.0280 vs. 0.0281).

The one-step-ahead out-of-sample forecasts from the nonlinear VECM model (model 3) are reported in the column labeled 3 in Table 3.[6] We find that the nonlinear VECM does not improve the forecasts of *SP* relative to the univariate model. In this case, we also find that the nonlinear model does not help with forecasting *BY*. It is interesting to note that the nonlinear VECM forecasts *SP* less accurately than the linear VECM (model 2), and the same result holds for *BY* as well.

The forecasts of the VAR in first differences (model 4) are better than those of the first three specifications for *SP* in terms of the RMSE measure, but are mixed with respect to the forecasting of *BY*.

The last column of Table 3 presents the forecast summary from the flexible linear specification in which we allow the cointegration rank to vary over recursive samples. Similar to the results under the first four model specifications, the predictability of *SP* is never better than when using the univariate model. In contrast, there seems to be some weak evidence that the inclusion of stock market information can improve the forecasts of the bond yield.[7]

4.3. Comparisons of Long-Run Forecasts

The above results on one-step-ahead forecasts may not be surprising, as the hypothesis of equal yield on equity and bonds is more likely to hold in the long run and hence may be more likely to improve longer-horizon forecasts. As argued in the literature, the use of long-run restrictions in a forecasting model – even when theoretically based restrictions are rejected in parametric tests – might well be more important for long-horizon forecasting performance.

Turning our attention to longer-horizon forecasts, Table 4 summarizes the performance of the six models for forecasting SP and BY 12-steps (months) ahead. As in the one-step-ahead forecasts, a simple VAR in first differences (model 4) outperforms the VECM models with estimated cointegrating vectors (models 1 and 5). In contrast, the second linear VECM model, where we impose the equal yield restriction in the cointegrating vector, provides better forecasts than the univariate model by both performance measures. We note, however, that while the performance differences are significant at the 0.10 significance level, the simulated critical values in McCracken (2004) are obtained assuming $h = 1$. Hence, they can be used only as rough guides in evaluating the long-run forecasts.

Note that when the nonlinearity is incorporated in the multivariate regression (in addition to the imposition of the equal yield restriction), the performance of the VECM model (model 3) further improves. It beats the univariate model now by relatively large margins.

The bottom panel of Table 4 shows that several VECM specifications generate better long-run forecasts than the univariate rival does. An exception is model 2, with the imposed cointegrating vector, at least when performance is measured by RMSE, and model 4, which has nearly the same RMSE but a higher MAE.

Figure 3 plots performance of the nonlinear VECM model (relative to the univariate model) against the out-of-sample forecasting horizon. Clearly, the performance of the nonlinear VECM model improves as the forecasting horizon increases (almost uniformly). For example, the RMSE of the nonlinear model is 1.05 times as large as that of the univariate model in one-step-ahead forecasts. When the horizon increases to six months, the two models perform equally well. At a 12-month horizon, the nonlinear model leads the competition by a ratio of 0.93:1.

The above result suggests that imposing restriction from theory may improve the forecasting performance of the error correction model, a finding consistent with some forecasting literature (e.g. Wang & Bessler, 2002). The result also indicates that the finding in previous studies that VECM models do not always forecast better than the unrestricted VAR model or univariate models, may be in part due to the parameter uncertainty in estimated cointegrating vectors and/or omitted nonlinearity.

These results may not be surprising in light of other results in the literature. Clements and Hendry (1995) and Hoffman and Rasche (1996) report that the inclusion of error correction terms in the forecasting model typically improves forecasts mostly at longer horizons. Rapach and Wohar (2005) also report such a result, and further report that a nonlinear ESTAR model fits the data

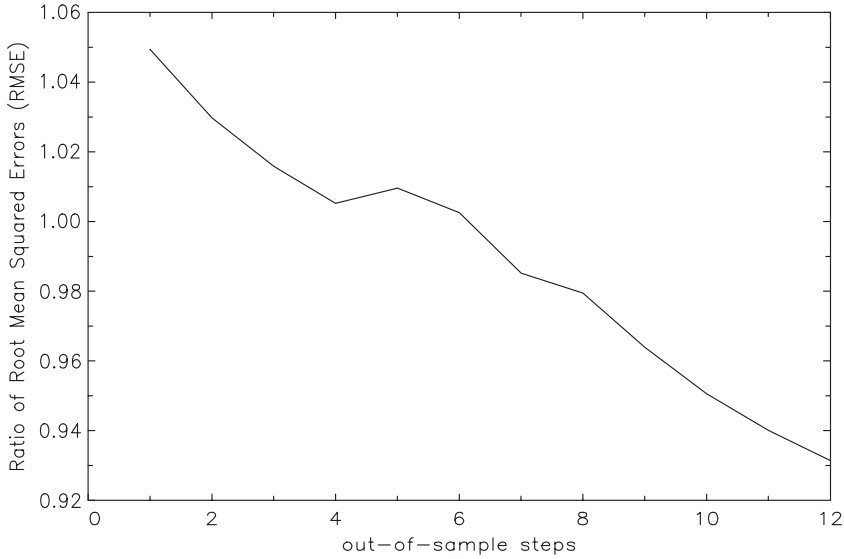


Fig. 3. The Ratio of Root Mean Squared Errors (RMSE) of Stock Price Forecasts from Unrestricted and Restricted ESTAR Models.

reasonably well and provides a plausible explanation for long-horizon stock price predictability, especially predictability based on the price-dividend ratio.

4.4. The Impact of the High-Tech Bubble on the Predictability of Stock Returns

In this subsection, we offer a brief discussion about how the stock price predictability may have been affected by the stock market boom in high-tech industry. As was widely observed, the period of the late 1990s featured historically low earning-to-price ratios, especially for the so-called high-tech stocks. These historically low earnings-to-price ratios may have had a large impact on the fit of our models. To provide some perspective, in Fig. 4 we plot the recursive estimates of RMSE statistics for both the nonlinear VECM and the univariate model for the forecasting period. We first focus on Panel A, which includes the one-step-ahead forecasts. In forecasting *SP*, the VECM performs better than the univariate model in the early period, but its performance deteriorates and is worse than the univariate model starting in late 1997. The VECM then continues to underperform the univariate model through the end of our sample.

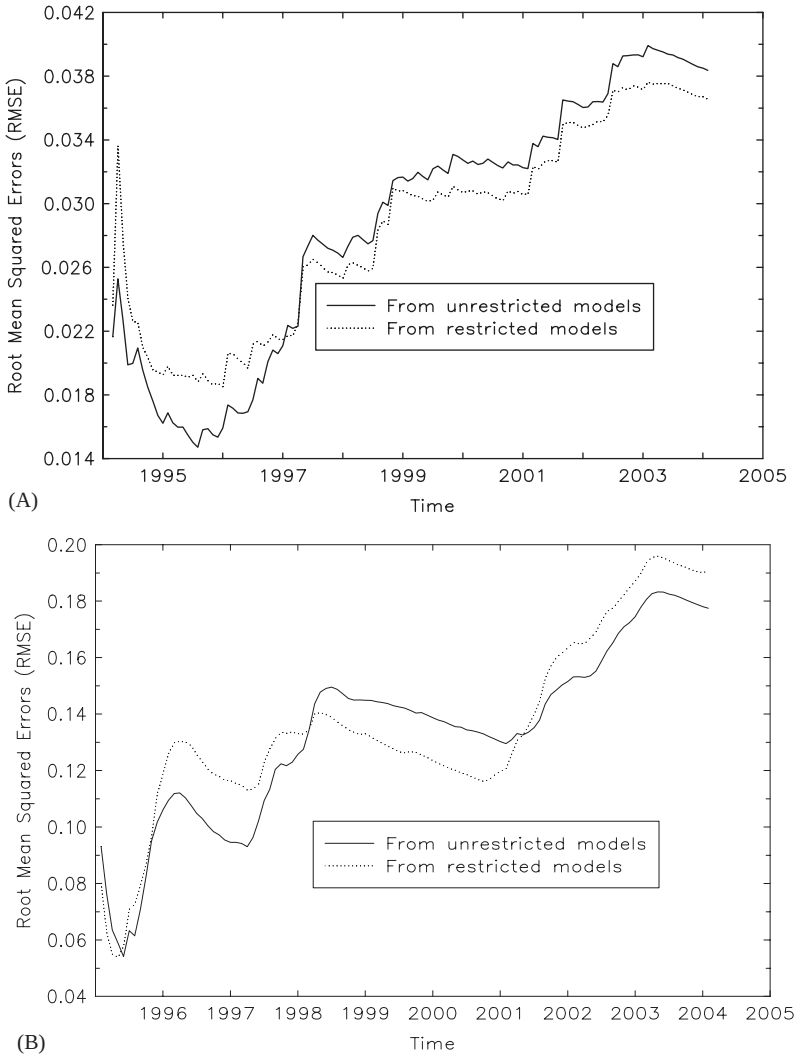


Fig. 4. Recursive Root Mean Squared Errors (RMSE) of Stock Prices. Panel (A) One-Step-Ahead Forecasts. Panel (B) 12-Step-Ahead Forecasts. *Note:* For the One-Step-Ahead Forecasts, the Statistics are Computed for the Period 1994M03–2004M02. For the 12-Step-Ahead Forecasts, the Statistics are Computed for the Period 1995M02–2004M02.

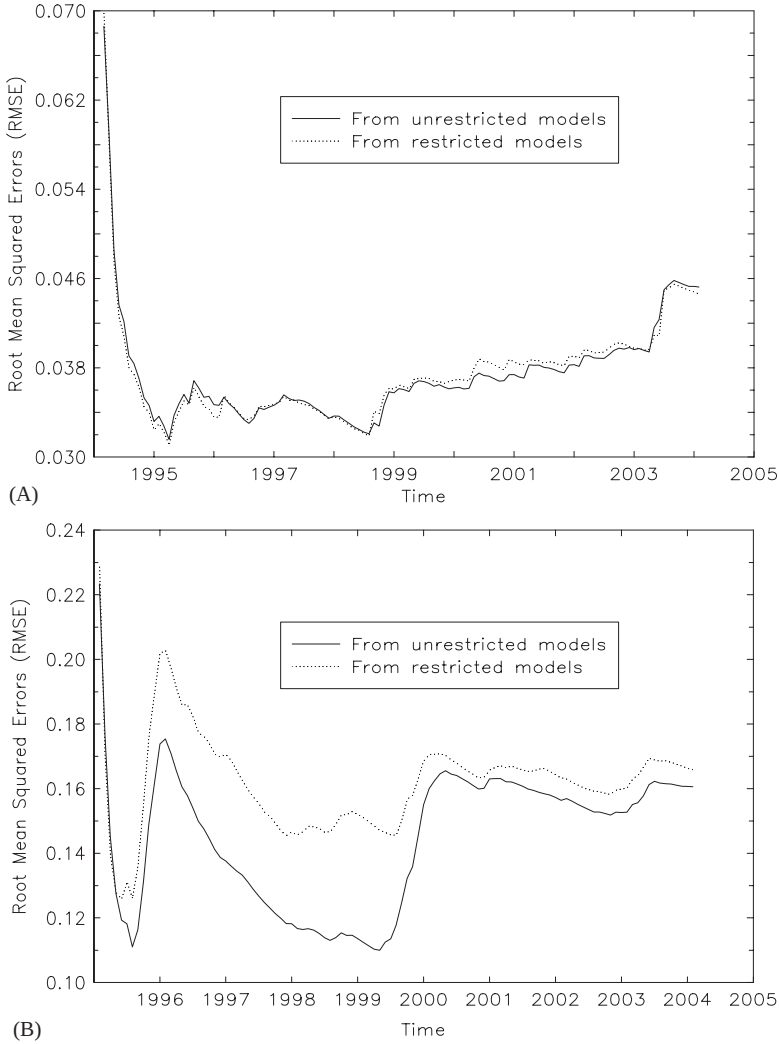


Fig. 5. Recursive Root Mean Squared Errors (RMSE) of Bond Yield. (A) One-Step-Ahead Forecasts. (B) 12-Step-Ahead Forecasts. *Note:* For the One-Step-Ahead Forecasts, the Statistics are Computed for the Period 1994M03–2004M02. For the 12-Step-Ahead Forecasts, the Statistics are Computed for the Period 1995M02–2004M02.

Panel B of Fig. 4 shows that the period of the high-tech boom was a period in which the longer-term predictability of stock returns also declined. After the stock market bust in 2001, information on *BY* again became helpful in long-run forecasts of *SP*, and the VECM forecasts were again better than the univariate model. Figure 5 plots the recursive RMSE

Table 5. Out-of-Sample Forecasts using Actual Earnings (Sample Period: 1994.03–2004.02).

Panel A: One-Step Ahead Forecasts			
	0	1	2
Forecasted variable: stock price (SP)			
Tested model	Univariate AR model	Linear VECM; Imposed CV	Nonlinear VECM
RMSE	0.0366	0.0374	0.0390
MAE	0.0281	0.0288	0.0293
D–M statistic	—	−0.969*	−1.556**
Forecasted variable: bond yield (BY)			
Tested model		Linear VECM; Imposed CV	Nonlinear VECM
RMSE	0.0446	0.0451	0.0452
MAE	0.0347	0.0351	0.0359
D-M statistic	—	−0.399	−1.137**
Panel B: 12-Month-Ahead Forecasts			
	0	1	2
Forecasted variable: stock price (SP)			
Tested model		Linear VECM; Imposed CV	Nonlinear VECM
RMSE	0.1906	0.1742	0.1653
MAE	0.1540	0.1476	0.1440
D–M statistic		0.759*	0.777**
Forecasted variable: bond yield (BY)			
Tested model	Linear VECM; Estimated CV	Linear VECM; Imposed CV	Nonlinear VECM
RMSE	0.1659	0.1915	0.2056
MAE	0.1380	0.1570	0.1708
D–M statistic		−1.789**	−2.328**

Note: Model 0 is a univariate AR Model; the comparison null model in each D–M test. Model 1 is a VECM with estimated cointegrating vector. Model 2 is a VECM with imposed cointegrating vector, EC = BY−SY.
**significance at the 0.05 level;
*significance at the 0.10 level.

estimates of BY for both one- and 12-step-ahead forecasts. It can be seen that the VECM consistently performs better than the univariate model in the long-run forecasts over the whole forecasting period.

Some may wonder if the results we have reported are due to our use of forecasted earnings. To investigate this possibility we reran our models using actual or reported earnings on the S&P 500. The results are similar, and are reported in Table 5. For the one-month-ahead forecasts of both SP and BY, the linear univariate models perform best. For the 12-month-ahead forecasts of SP, the linear VECM model with imposed cointegrating vector performs better than the univariate model, and the nonlinear VECM performs even better than the linear VECM. For 12-month ahead predictions of BY, however, the univariate model outperforms either variant of VECM.

We summarize the results relating forecasting horizon to forecasting performance for forecasts of SP in Fig. 6. Similarly to our results with forward earnings, we find that the performance gain of the nonlinear VECM increases nearly monotonically with forecast horizon beginning at the five-month horizon.

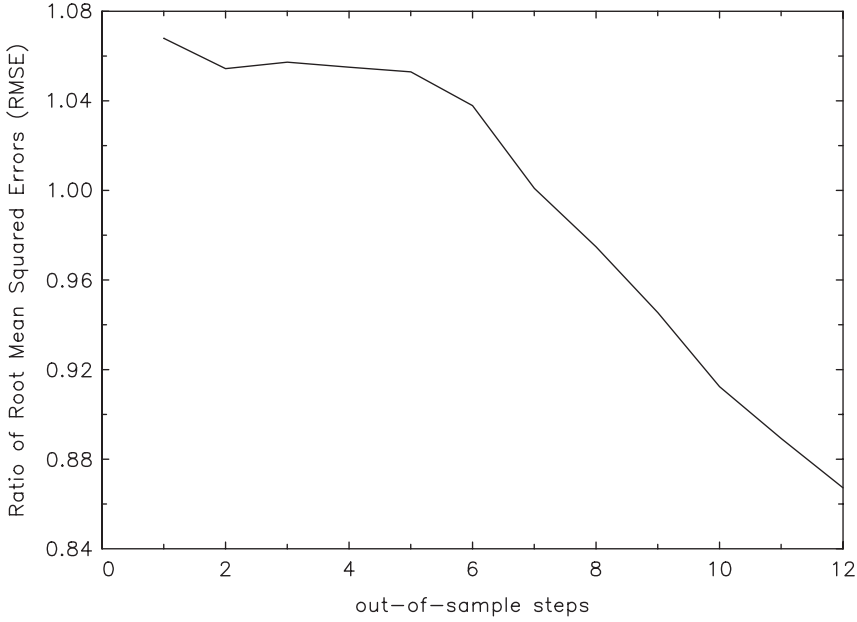


Fig. 6. The Ratio of Root Mean Squared Errors (RMSE) of Stock Price Forecasts from Unrestricted and Restricted ESTAR Models using Actual Earnings.

4.5. Nonlinear Dynamics in the Stock Market

Finally, we provide some details about the nonlinear VECM model. Table 6 summarizes the parameter estimates from the nonlinear VECM based on the full sample. (We do not report the estimates of the linear VECM and the univariate model, but they are available from the authors upon request.) Of interest is the coefficient θ, which indicates the curvature of the transition function $F(\cdot)$. The estimate is moderate in magnitudes in both SP and BY equations, which is easier to see in Fig. 7 where we plot the estimated transition functions. Most observations of EC variable (delay variable) fall in the range $-0.2 - 0.2$. Within this range, the transition function has values from 0 to 0.5. Note that the limiting case of $F = 1$ is not attained over the range of observed yield differences, as the maximum difference is 0.543 with a transition function value of 0.987 (the observed minimum deviation is –0.491 with a F-value of 0.971). Panel B graphs the transition function for the BY equation, which essentially tells the same story. Finally, Panel C graphs the transition function for the FE equation, which has a much higher value for θ and hence a much narrower range in which the transition function is near zero.

Table 6. Parameter Estimates from the STAR Model, 1979.01 – 2004.02.

Dependent variable						
	Stock Price (SP)		Bond Yield (BY)		Forward Earnings (FE)	
	Parameter est.	*p*-value	Parameter est.	*p*-value	Parameter est.	*p*-value
μ	0.0056	0.0157	−0.0039	0.1191	0.0037	0.0001
ΔFE_{t-1}	0.0495	0.7945	−0.1307	0.5862	0.4816	0.0004
ΔSP_{t-1}	0.1928	0.0090	0.1032	0.2835	0.1155	0.0026
ΔBY_{t-1}	−0.1235	0.0868	0.4260	0.0000	0.1163	0.0023
EC_{t-1}	0.0880	0.0128	0.0036	0.9252	0.0015	0.8545
$\Delta FE_{t-1} * F(EC_{t-d})$	−0.1010	0.8124	0.6959	0.1181	−0.4403	0.0248
$\Delta SP_{t-1} * F(EC_{t-d})$	−0.1436	0.4608	0.2742	0.1508	−0.1636	0.0119
$\Delta BY_{t-1} * F(EC_{t-d})$	0.1097	0.4482	−0.2764	0.0801	−0.1312	0.0168
$EC_{t-1} * F(EC_{t-d})$	−0.0728	0.0823	0.0193	0.6668	0.0126	0.2592
θ	14.6483	0.2284	22.5343	0.1217	13.9508	0.0231
Mean log-likelihood	1.883		1.804		2.9566	

Note: The error correction term is $EC_{t-1} = FE_{t-1} - SP_{t-1} - BY_{t-1}$. The transition function is exponential, the transition variable is EC_{t-d} where $d = 1$ for the SP equation, and $d = 3$ for the BY equation. The transition function is logistic, the transition variable is EC_{t-d} where $d = 3$.

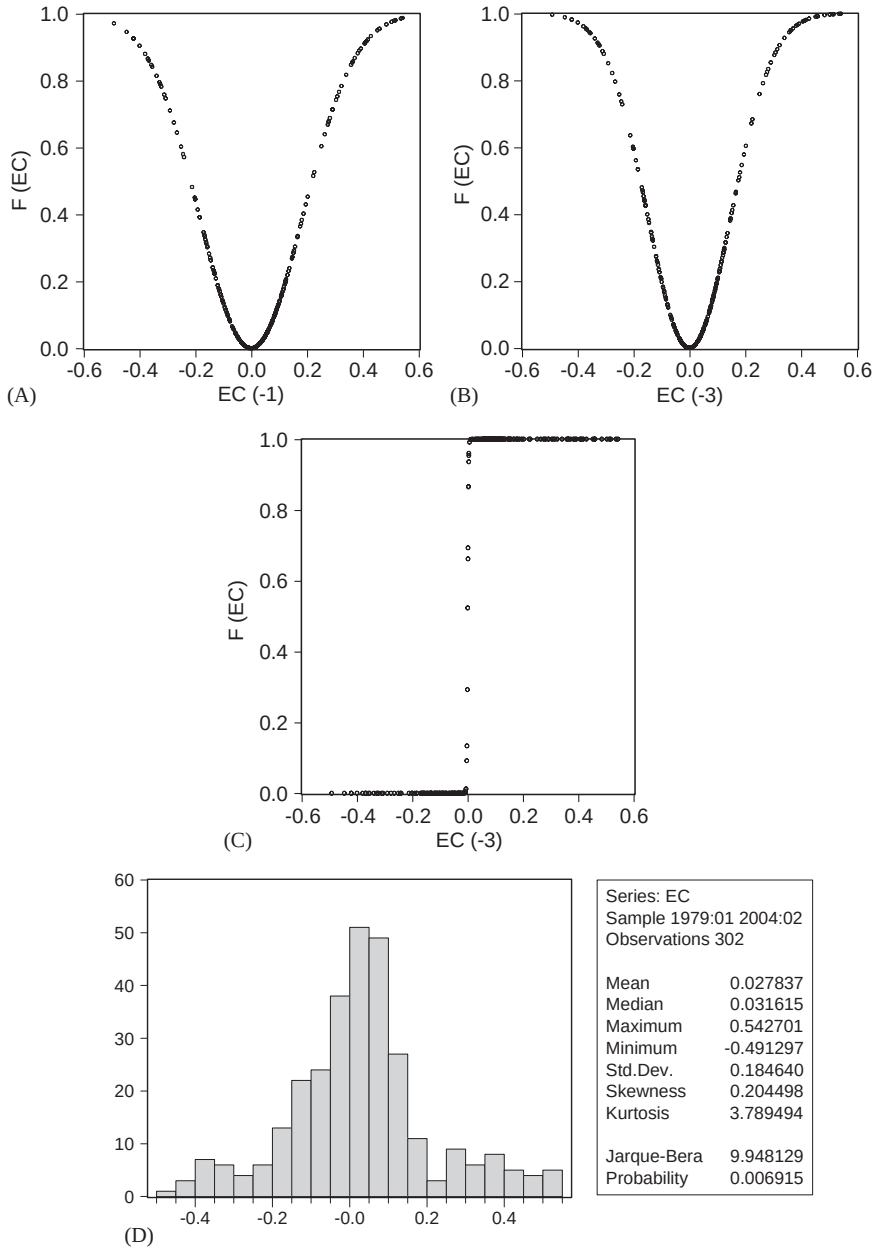


Fig. 7. Transition Functions. Panel (A) Stock Price. Panel (B) Bond Yield. Panel (C) Forward Earnings. Panel (D) Histogram of EC.

5. CONCLUSION

The Fed Model of stock market valuation postulates a cointegrating relationship between the equity yield on the S&P 500 (earnings over index value) and the bond yield. We evaluate the Fed Model as a vector error correction forecasting model for stock prices and for bond yields, and compare out-of-sample forecasts of each of these two variables from the Fed Model to forecasts from alternative models that include a univariate model, a nonlinear vector error correction model, and a set of modified vector error correction models. We find that the Fed Model does not improve on the forecasts of the univariate models at short horizons, but that the Fed Model is able to improve on the univariate model for longer-horizon forecasts, especially for stock prices. The nonlinear vector error correction model performs even better than its linear version as a forecasting model for stock prices, although this does not hold true for forecasts of the bond yield.

NOTES

1. The STAR model has been widely used in finance and economics literature (e.g. Terasvirta & Anderson, 1992; Anderson, 1997; Michael, Nobey, & Peel, 1997; Jansen and Oh, 1999; Bradley & Jansen, 2004).

2. We note that in the full sample information SIC attain its minimum at $p = 2$, while the choice of AIC remains $p = 4$.

3. We examined the models with $p = 4$ and found that the forecasts of SP were less accurate than those from the corresponding models with $p = 2$. In contrast, we found that forecasts of FE were more accurate in the VAR(4). As our focus is the predictability of stock prices, and the basic conclusions in the paper are not changed by using other p-values (i.e. $p = 1, 3, 4$), we only report the results associated with $p = 2$.

4. We also conducted a second exercise in which we sequentially test the cointegrating rank using 121 rolling fixed-window samples. The first sample is the period March 1979–February 1994, 180 months, and we continue changing the starting and ending period by one month until we reach our final sample period of March 1994–February 2004. This procedure allows for a quicker detection of underlying structural breaks at the cost of never having a sample size that grows larger over time with the addition of new observations. The rolling sample estimates are consistent with Panel A but if anything indicate an earlier breakdown of the cointegrating relationship, in 1997.

5. The p-value of the null hypothesis of equal yield for the full sample is 0.05.

6. Note that the in-sample linearity test suggests that the delay parameter d is 1 for the BY equation. However, we find that, while basic results to be reported below remain the same for SP for any d-values, the specification that $d = 3$ for the BY equation provides somewhat better out-of-sample forecasts. This is consistent with the linearity test based on the full sample, where we find that the test obtains the

minimum p-value at $d = 3$ (the p-values are 0.068, 0.207 and 0.037 for $d = 1$, 2 and 3, respectively). This may be due to the stock market responding more quickly to innovations than the bond. In any case, we report results for $d = 3$ for the BY equation throughout the paper.

7. We also investigated the performance of a rolling fixed-window linear VECM. Results were similar to the recursive model reported in Table 3.

REFERENCES

Anderson, H. (1997). Transition costs and non-linear adjustment toward equilibrium in the U.S. Treasury bill market. *Oxford Bulletin of Economics and Statistics*, *59*, 465–484.

Asness, C. (2003). Fight the Fed Model: The relationship between future returns and stock and bond market yields. *Journal of Portfolio Management*, *30*, 11–24.

Bradley, M. D., & Jansen, D. W. (2004). Forecasting with a nonlinear dynamic model of stock returns and industrial production. *International Journal of Forecasting*, *20*, 321–342.

Clark, T., & McCracken, M. W. (2004). *Evaluating long-horizon forecasts*. Manuscript, University of Missouri-Columbia.

Clements, M. P., & Hendry, D. F. (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics*, *10*, 127–146.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economics Statistics*, *13*, 253–263.

Hoffman, D. L., & Rasche, R. H. (1996). Assessing forecast performance in a cointegrated system. *Journal of Applied Econometrics*, *11*, 495–517.

Jansen, D. W., & Oh, W. (1999). Modeling nonlinearity of business cycles: Choosing between the CDR and STAR models. *Review of Economics and Statistics*, *81*, 344–349.

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.

Kilian, L., & Taylor, M. P. (2003). Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, *14*, 85–107.

McCracken, M. W. (2004). *Asymptotics for out-of-sample tests of granger causality*. Working Paper, University of Missouri-Columbia.

Michael, P. A., Nobay, R., & Peel, D. A. (1997). Transaction costs and nonlinear adjustment in real exchange rates: An empirical investigation. *Journal of Political Economy*, *105*, 862–879.

Rapach, D. E., & Wohar, M. E. (2005). Valuation ratios and long-horizon stock price predictability. *Journal of Applied Econometrics*, *20*, 327–344.

Seo, B. (1996). Tests for structural change in cointegrated systems. *Econometric Theory*, *14*, 222–259.

Terasvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of American Statistical Association*, *89*, 208–218.

Terasvirta, T., & Anderson, H. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, *7*, S119–S136.

Wang, Z., & Bessler, D. A. (2002). The homogeneity restriction and forecasting performance of VAR-type demand systems: An empirical examination of U.S. meat consumption. *Journal of Forecasting*, *21*, 193–206.

STRUCTURAL CHANGE AS AN ALTERNATIVE TO LONG MEMORY IN FINANCIAL TIME SERIES

Tze Leung Lai and Haipeng Xing

ABSTRACT

This paper shows that volatility persistence in GARCH models and spurious long memory in autoregressive models may arise if the possibility of structural changes is not incorporated in the time series model. It also describes a tractable hidden Markov model (HMM) in which the regression parameters and error variances may undergo abrupt changes at unknown time points, while staying constant between adjacent changepoints. Applications to real and simulated financial time series are given to illustrate the issues and methods.

1. INTRODUCTION

Volatility modeling is a cornerstone of empirical finance, as portfolio theory, asset pricing and hedging all involve volatilities. Since the seminal works of Engle (1982) and Bollerslev (1986), generalized autoregressive conditionally heteroskedastic (GARCH) models have been widely used to model and forecast volatilities of financial time series. In many empirical studies of

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 205–224
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20027-0

stock returns and exchange rates, estimation of the parameters ω, α and β in the GARCH(1,1) model

$$y_n = \sigma_n \varepsilon_n, \quad \sigma_n^2 = \omega + \alpha y_{n-1}^2 + \beta \sigma_{n-1}^2 \tag{1}$$

reveals high volatility persistence, with the maximum likelihood estimate $\hat{\lambda} = \hat{\alpha} + \hat{\beta}$ close to 1. The constraint $\lambda < 1$ in the GARCH model (1), in which the ε_n are independent standard normal random variables that are not observable and the y_n are the observations, enables the conditional variance σ_n^2 of y_n given the past observations to be expressible as

$$\sigma_n^2 = (1-\lambda)^{-1}\omega + \alpha\{(y_{n-1}^2 - \sigma_{n-1}^2) + \lambda(y_{n-2}^2 - \sigma_{n-2}^2) + \lambda^2(y_{n-3}^2 - \sigma_{n-3}^2) + \cdots\}$$

For $\lambda = 1$, the contributions of the past innovations $y_{n-t}^2 - \sigma_{n-t}^2$ to the conditional variance do not decay over time but are "integrated" instead, yielding the IGARCH model of Engle and Bollerslev (1986). Baillie, Bollerslev, and Mikkelsen (1996) introduced fractional integration in their FIGARCH models, with a slow hyperbolic rate of decay for the influence of the past innovations, to quantify the long memory of exchange rate volatilities.

In his comments on Engle and Bollerslev (1986), Diebold (1986) noted with respect to interest rate data that the choice of a constant term ω in (1) not accommodating to shifts in monetary policy regimes might have led to an apparently integrated series of innovations. By assuming ω to be piecewise constant and allowing the possibility of jumps between evenly spaced time intervals, Lamoureux and Lastraps (1990) obtained smaller estimates of λ, from the daily returns of 30 stocks during the period 1963–1979, than those showing strong persistence based on the usual GARCH model with constant ω.

In Section 2 we carry out another empirical study of volatility persistence in the weekly returns of the NASDAQ index by updating the estimates of the parameters ω, α and β in the GARCH model (1) as new data arrive during the period January 1986 to September 2003, starting with an initial estimate based on the period November 1984 to December 1985. These sequential estimates show that the parameters changed over time and that $\hat{\lambda}_N$ eventually became very close to 1 as the sample size N increased with accumulating data. The empirical study therefore suggests using a model that incorporates the possibility of structural changes for these data. In Section 3 we describe a structural change model, which allows changes in the volatility and regression parameters at unknown times and with unknown change in magnitudes. It is a hidden Markov model (HMM), for which the

volatility and regression parameters at time t based on data up to time t (or up to time $n > t$) can be estimated by recursive filters (or smoothers).

Although we have been concerned with volatility persistence in GARCH models so far, the issue of spurious long memory when the possibility of structural change is not incorporated into the time series model also arises in autoregressive and other models. In Section 4, by making use of the stationary distribution of a variant of the structural change model in Section 3, we compute the asymptotic properties of the least squares estimates in a nominal AR(1) model that assumes constant parameters, thereby showing near unit root behavior of the estimated autoregressive parameter. Section 5 gives some concluding remarks and discusses related literature on structural breaks and the advantages of our approach.

2. VOLATILITY PERSISTENCE IN NASDAQ WEEKLY RETURNS

Figure 1, top panel, plots the weekly returns of the NASDAQ index, from the week starting on November 19, 1984 to the week starting on September 15, 2003. The series r_t is constructed from the closing price P_t on the last day of the week via $r_t = 100 \log(P_t/P_{t-1})$. The data consisting of 982 observations are available at http://finance.yahoo.com. Similar to Lamoureux and Lastraps (1990, p. 227), who use lagged dependent variables to account for nonsynchronous trading, we fit an AR(2) model $r_n = \mu + \rho_1 r_{n-1} + \rho_2 r_{n-2} + y_n$ to the return series. The y_n are assumed to follow a GARCH(1,1) process so that the ε_n in (1) is standard normal and independent of $\{(\varepsilon_i,\ y_i,\ r_i),\quad i \le n-1\}$. The parameters of this AR-GARCH model can be estimated by maximum likelihood using 'garchfit' in MATLAB. Besides using the full dataset, we also estimated these parameters sequentially from January 1986 to September 2003, starting with an initial estimate based on the period November 1984 to December 1985. These sequential estimates are plotted in Fig. 2. Note that the estimates on the last date, September 15, 2003, corresponds to those based on the full dataset; see the last row of Table 1.

Lamoureux and Lastraps (1990) proposed the following strategy to incorporate possible time variations in ω in fitting GARCH(1,1) models to a time series of $N = 4,228$ observations of daily returns of each of 30 stocks during the period 1963–1979. They allow for possible jumps in ω every 302 observations, thereby replacing ω in (1) by $\omega' + \delta_1 D_{1n} + \cdots + \delta_k D_{kn}$, where $D_{1n}, \ldots, D_{kn}$ are dummy variables indicating the subsample to which n belongs. The choice of 302, 604, 906, … as the only possible jump points of ω

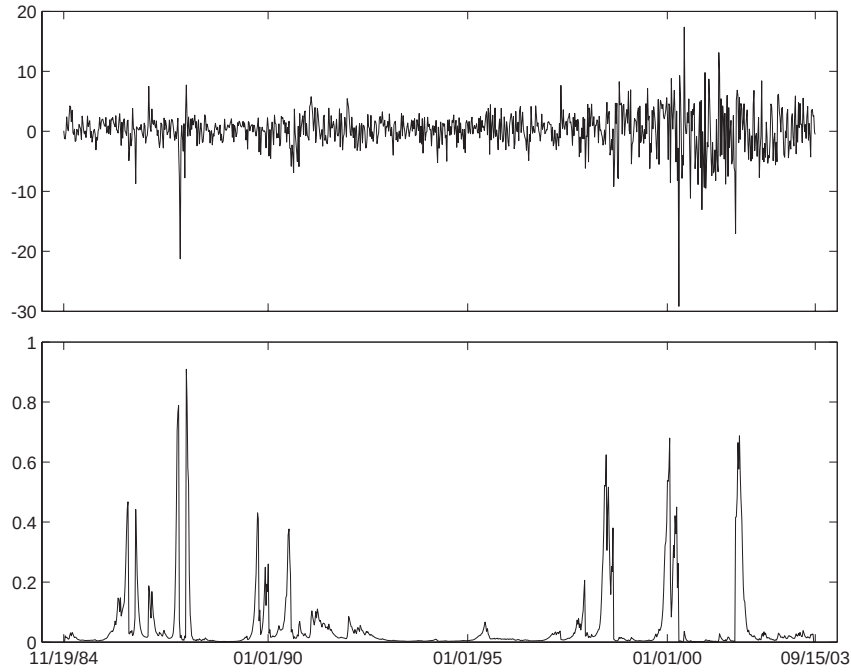


Fig. 1. Weekly Returns (Top Panel) of NASDAQ Index and Estimated Probabilities of Structural Change (Bottom Panel).

while not allowing jumps in α and β appears to be overly restrictive. In Section 3.3 we fit a HMM of structural change to the observed time series and use the fitted model to estimate the probability of changes in both the autoregressive and volatility parameters at any time point. On the basis of these estimated probabilities, which are plotted in the bottom panel of Fig. 1, we divide the time series of NASDAQ index returns into 4 segments of different lengths; see Section 3.3 for further details. The estimates of the AR-GARCH parameters in each segment are compared with those based on the entire dataset in Table 1. Consecutive segments in Table 1 have slight overlap because we try not to initialize the returns with highly volatile data for a segment. Table 1 shows substantial differences in the estimated AR and GARCH parameters among the different time periods. In particular, $\hat{\lambda}$ is as small as 0.0364 in the first segment but as high as 0.9858 in the last segment. Although this seems to suggest volatility persistence during the last period of $5\frac{1}{2}$ years, we give an alternative viewpoint at the end of this Section.

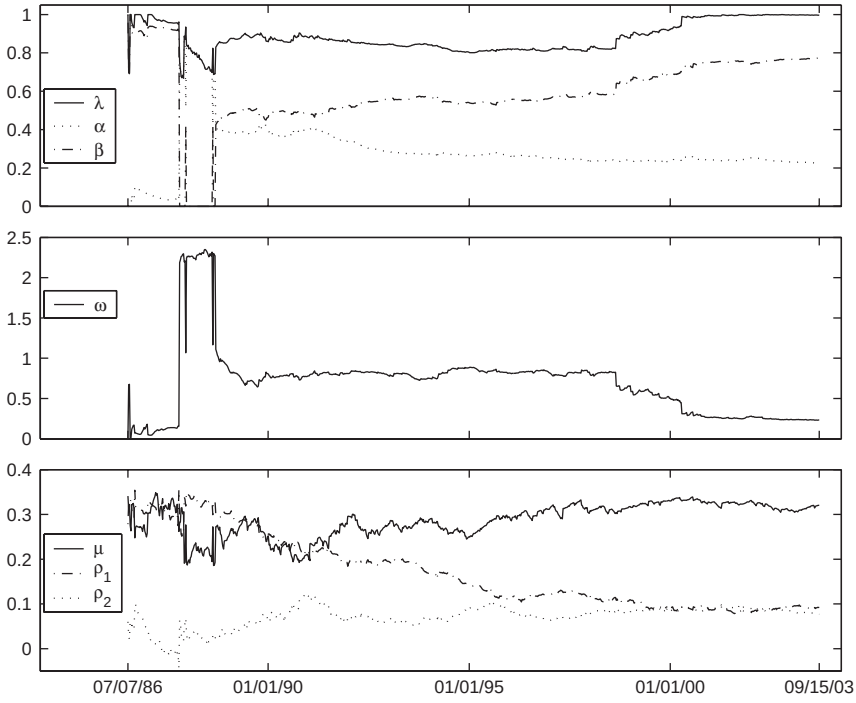


Fig. 2. Sequential Estimates of AR-GARCH Parameters.

Table 1. Estimated AR(2)-GARCH(1,1) Parameters for Different Periods.

Period	μ	ρ_1	ρ_2	ω	α	β	λ
Nov 19, 1984–Jun 16, 1987	0.3328	0.3091	0.09838	1.7453	0.0364	0.0	0.0364
Jun 09, 1987–Aug 15, 1990	0.1242	0.2068	0.0891	0.8364	0.4645	0.4784	0.9429
Aug 08, 1990–Mar 13, 1998	0.3362	0.03874	0.1039	0.4560	0.08767	0.8008	0.8885
Mar 06, 1998–Sep 15, 2003	0.2785	0.05612	0.04063	0.8777	0.2039	0.7819	0.9858
Nov 19, 1984–Sep 15, 2003	0.3203	0.09241	0.07657	0.2335	0.2243	0.7720	0.9963

Using the estimates $\hat{\mu}$ of μ, $\hat{\rho}_1$ and $\hat{\rho}_2$ of the autoregressive parameters and $\hat{\omega}$, $\hat{\alpha}$ and $\hat{\beta}$ of the GARCH parameters, we can compute the estimated mean return $\hat{\mu} + \hat{\rho}_1 r_{n-1} + \rho_2 \hat{r}_{n-2}$ at time n, the residual $\hat{y}_n = r_n - (\hat{\mu} + \hat{\rho}_1 r_{n-1} + \hat{\rho}_2 r_{n-2})$, and the volatility $\hat{\sigma}_n = (\hat{\omega} + \hat{\alpha}\hat{y}_{n-1}^2 + \hat{\beta}\,\hat{\sigma}_{n-1}^2)^{1/2}$ recursively. The

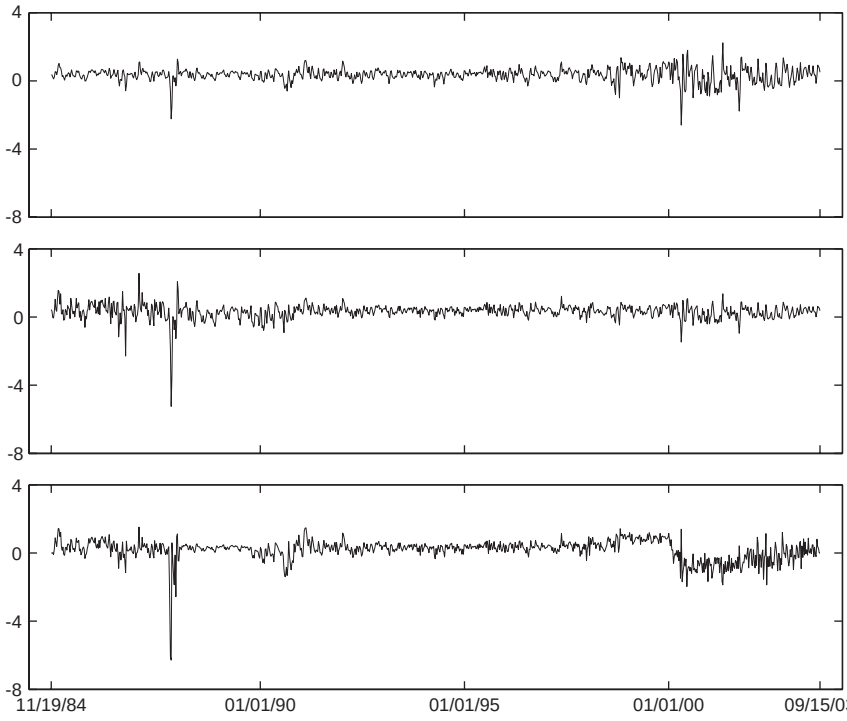


Fig. 3. Mean Return Estimates Using Three Different Methods.

results are plotted in Figs. 3 and 4, with the top panel using parameter estimates based on the entire dataset and the middle panel using different parameter estimates for different segments of the data. The bottom panel of Fig. 3 plots the estimated mean returns based on a HMM of structural change (see Section 3.2). The bottom panel of Fig. 4 plots (1) the estimated volatilities (solid curve) based on the structural change model (see Section 3.2), and (2) standard deviations (dotted curve) of the residuals in fitting an AR(2) model to the current and previous 19 observations (see Section 3.3 for further details).

Consider the last $5\frac{1}{2}$ year period in Table 1 that gives $\hat{\lambda} = 0.9858$. An alternative to the GARCH model, which is used for the bottom panel of Fig. 4, is the structural change model, which assumes that $y_n = \sigma_n \varepsilon_n$ with σ_n undergoing periodic jumps according to a renewal process, unlike the continuously changing σ_n in the GARCH model (1). As the σ_n are unknown, we replace them by their estimates $\hat{\sigma}_n$ (given by the solid curve

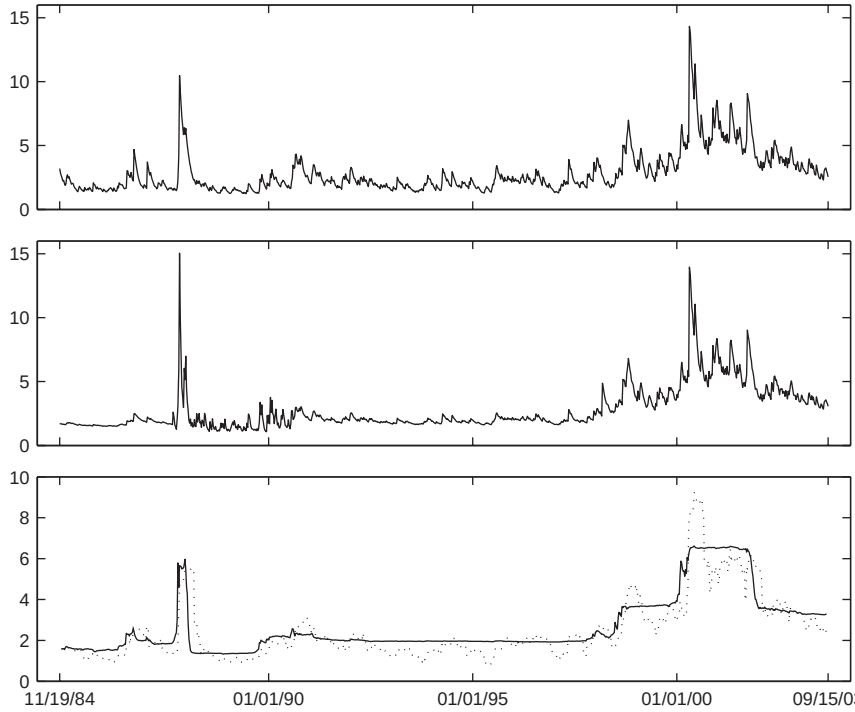


Fig. 4. Volatility Estimates Using Three Different Methods.

in the bottom panel of Fig. 4) and consider a simulated set of innovations $y_n^* = \hat{\sigma}_n \varepsilon_n^*$ for this last segment of the data, in which ε_n^* are i.i.d. standard normal random variables. Fitting a nominal GARCH (1,1) model to these simulated innovations yielded a value $\hat{\lambda} = 0.9843$, which is also close to 1, and differs little from value of 0.9858 for the last period in Table 1.

3. AUTOREGRESSIVE MODELS WITH STRUCTURAL BREAKS IN VOLATILITY AND REGRESSION PARAMETERS

Bayesian modeling of structural breaks in autoregressive time series has been an active research topic in the last two decades. Hamilton (1989) proposed a regime-switching model, which treats the autoregressive parameters as the hidden states of a finite-state Markov chain. Noting that volatility

persistence of stock returns may be overestimated when one ignores possible structural changes in the time series of stock returns, Hamilton and Susmel (1994) extended regime switching for autoregression to a Markov-switching ARCH model, which allows the parameters of Engle's (1982) ARCH model of stock returns to oscillate among a few unspecified regimes, with transitions between regimes governed by a Markov chain. Albert and Chib (1993) considered autoregressive models with exogenous variables whose autoregressive parameters and error variances are subject to regime shifts determined by a two-state Markov chain with unknown transition probabilities. Wang and Zivot (2000) introduced a Bayesian time series model which assumes the number of the change-points to be known and in which multiple change points in level, trend and error variance are modeled using conjugate priors. Similar Bayesian autoregressive models allowing structural changes were introduced by Carlin, Gelfand, and Smith (1992), McCulloch and Tsay (1993), Chib (1998) and Chib, Nardari, and Shepard (2002). However, except for Hamilton's (1989) seminal regime-switching model, Markov Chain Monte Carlo techniques are needed in the statistical analysis of these models because of their analytic intractability.

Lai, Liu, and Xing (2005) recently introduced the following model to incorporate possible structural breaks in the AR(k) process

$$X_n = \mu_n + \rho_{1n} X_{n-1} + \cdots + \rho_{kn} X_{n-k} + \sigma_n \varepsilon_n, \quad n > k \tag{2}$$

where the ε_n are i.i.d. unobservable standard normal random variables, and $\boldsymbol{\theta}_n = (\mu_n,\ \rho_{1n}, \ldots, \rho_{kn})^T$ and σ_n are piecewise constant parameters. Note that σ_n^2 is assumed to be piecewise constant in this model, unlike the continuous change specified by the linear difference Eq. (1). The sequence of change-times of $(\boldsymbol{\theta}_t^T,\ \sigma_t)$ is assumed to form a discrete renewal process with parameter p, or equivalently,

$I_t := 1_{\{(\boldsymbol{\theta}_t,\ \sigma_t) \neq (\boldsymbol{\theta}_{t-1},\ \sigma_{t-1})\}}$ are i.i.d. Bernoulli random variables with $P(I_t = 1) = p$ for $t \geq k+2$ and $I_{k+1} = 1$. In addition, letting $\tau_t = (2\sigma_t^2)^{-1}$, it is assumed that

$$(\boldsymbol{\theta}_t^T, \tau_t) = (1 - I_t)\,(\boldsymbol{\theta}_{t-1}^T,\ \tau_{t-1}) + I_t (z_t^T, \gamma_t) \tag{3}$$

where $(\mathbf{Z}_1^T,\ \gamma_1)$, $(\mathbf{Z}_2^T, \gamma_2)$, ... are i.i.d. random vectors such that,

$$\gamma_t \sim \text{Gamma}\,(g,\ \lambda), \quad \mathbf{Z}_t | \gamma_t \sim \text{Normal}\,(\text{z},\ \mathbf{V}/(2\gamma_t)) \tag{4}$$

An important advantage of this Bayesian model of structral breaks is its analytic and computational tractability. Explicit recursive formulas are available for estimating σ_n, μ_n, $\rho_{1,n}$,... and $\rho_{k,n}$ in the Bayesian model (2)–(4).

As we now proceed to show, these formulas also have an intuitively appealing interpretation as certain weighted averages of the estimates based on different segments of the data and assuming that there is no change-point in each segment.

3.1. Sequential Estimation of σ_n, μ_n, $\rho_{1,n}, \ldots, \rho_{k,n}$ via Recursive Filters

To estimate $(\boldsymbol{\theta}_n^T, \sigma_n^2)$ from current and past observations $X_1, \ldots, X_n$, let $\mathbf{X}_{t,n} = (1, X_n, \ldots, X_t)^T$ and consider the most recent change-time $J_n := \max\{t \le n : I_t = 1\}$. Recalling that $\tau_n = (2\sigma_n^2)^{-1}$, the conditional distribution of $(\boldsymbol{\theta}_n^T, \tau_n)$ given $(J_n, \mathbf{X}_{J_n}, n)$ can be described by

$$\tau_n \sim \text{Gamma}\left(g + \frac{n - J_n + 1}{2}, \frac{1}{a_{J_n,n}}\right), \ \boldsymbol{\theta}_n | \tau_n \sim \text{Normal}\left(\mathbf{z}_{J_n,n}, \frac{1}{2\tau_n}\mathbf{V}_{J_n,n}\right) \tag{5}$$

where for $k < j \le n$,

$$\begin{aligned} \mathbf{V}_{j,n} &= \left(\mathbf{V}^{-1} + \sum_{t=1}^{n} \mathbf{X}_{t-k,\,t-1}\mathbf{X}_{t-k,\,t-1}^T\right)^{-1}, \ \mathbf{z}_{j,n} = \mathbf{V}_{j,n}\left(\mathbf{V}^{-1}\mathrm{z} + \sum_{\mathrm{t}=1}^{\mathrm{n}} \mathbf{X}_{t-k,\,t-1}X_t\right) \\ a_{j,n} &= \lambda^{-1} + \mathbf{z}^T\mathbf{V}^{-1}\mathbf{z} + \sum_{t=j}^{n} X_t^2 - \mathbf{z}_{j,n}^T\mathbf{V}_{j,n}^{-1}\mathrm{z}_{j,n} \end{aligned} \tag{6}$$

Note that if $(2Y)^{-1}$ has a Gamma $(\tilde{g}, \tilde{\lambda})$ distribution, then Y has the inverse gamma IG(g,λ) distribution with $g = \tilde{g}$, $\lambda = 2\tilde{\lambda}$ and that $EY = \lambda^{-1}(g-1)^{-1}$ when $g > 1$ and $E\sqrt{Y} = \lambda^{-1/2}\Gamma(g - \frac{1}{2})/\Gamma(g)$. It then follows from (5) that

$$\begin{aligned} E(\boldsymbol{\theta}_n^T, \sigma_n^2 | \mathbf{X}_{1,n}) &= \sum_{j=k+1}^{n} p_{j,n} E(\boldsymbol{\theta}_n^T, \sigma_n^2 | \mathbf{X}_{1,n}, J_n = j) \\ &= \sum_{j=k+1}^{n} p_{j,n}\left(\mathbf{z}_{j,n}^T, \frac{a_{j,n}}{2g + n - j - 1}\right) \end{aligned} \tag{7}$$

where $p_{j,n} = P(J_n = j | \mathbf{X}_{1,n})$; see Lai et al. (2005, p. 282). Moreover, $E(\sigma_n | \mathbf{X}_{1,n}) = \Sigma_{j=k+1}^{n} p_{j,n}(\mathrm{a}_{j,n}/2)^{1/2}\Gamma_{n-j}$, where $\Gamma_i = \Gamma(g + i/2)\Gamma(g + (i+1)/2)$. Denoting conditional densities by $f(\cdot|\cdot)$, the weights $p_{j,n}$ can be determined recursively by

$$p_{j,n} \propto p_{j,n}^* := \begin{cases} pf(X_n | J_n = j) & \text{if } j = n \\ (1-p)p_{j,n-1} f(X_n | \mathbf{X}_{j-k,n-1}, J_n = j) & \text{if } j \le n-1 \end{cases} \tag{8}$$

Since $\sum_{i=k+1}^{n} p_{i,n} = 1$, $p_{j,n} = p^*_{j,n} / \sum_{i=k+1}^{n} p^*_{i,n}$. Moreover, as shown in Lemma 1 of Lai et al. (2005),

$$f(\lambda_j^{-1}(X_n - \mathbf{z}_{j,n-1}^T \mathbf{X}_{j-k,n-1}) | J_n = j,\ \mathbf{X}_{j-k,n-1}) = \text{Stud}(g_j) \tag{9}$$

where Stud(g) denotes the Student-t density function with g degrees of freedom and

$$\begin{aligned} &g_j = 2g + n - j, \quad && \lambda_j = a_{j,n}(1 + \mathbf{X}_{n-k+1,n-1}^T \mathbf{V} \mathbf{X}_{n-k+1,n-1})/(2g + n - j) && \text{if } j < n \\ &g_j = 2g, && \lambda_j = (1 + \mathbf{X}_{n-k+1,n-1}^T \mathbf{V} \mathbf{X}_{n-k+1,n-1})/(2\lambda g) && \text{if } j = n \end{aligned}$$

3.2. *Estimating Piecewise Constant Parameters from a Time Series via Bayes Smoothers*

Lai et al. (2005) evaluate the minimum variance estimate $E(\tau_t, \boldsymbol{\theta}_t^T | X_1, \ldots, X_n)$, with $k+1 \le t \le n$, by applying Bayes' theorem to combine the forward filter that involves the conditional distribution of $(\tau_t,\ \boldsymbol{\theta}_t^T)$ given $X_1, \ldots, X_t$ and the backward filter that involves the conditional distribution of $(\tau_t,\ \boldsymbol{\theta}_t^T)$ given $X_{t+1}, \ldots, X_n$. Noting that the normal distribution for $\boldsymbol{\theta}_t$ assigns positive probability to the explosive region $\{\boldsymbol{\theta} = (\mu, \rho_1, \ldots, \rho_k)^T : 1 - \rho_1 z - \cdots - \rho_k z^k\}$ has roots inside the unit circle, they propose to replace the normal distribution in (4) by a truncated normal distribution that has support in some stability region C such that

$$\inf_{|z| \le 1} |1 - \rho_1 z - \cdots - \rho_k z^k| > 0 \ \text{ if } \ \boldsymbol{\theta} = (\mu, \rho_1, \ldots, \rho_k)^T \in C \tag{10}$$

Letting $\mathrm{T}_C\,\text{Normal}(\mathbf{z}, \mathrm{V})$ denote the conditional distribution of Z given $\mathrm{Z} \in C$, where $\mathrm{Z} \sim \text{Normal}(\mathbf{z}, \mathrm{V})$ and C satisfies the stability condition (10), they modify (4) as

$$\gamma_t \sim \text{Gamma}\,(g, \lambda), \ \ \mathbf{Z}_t | \gamma_t \sim \mathbf{T}_C \text{Normal}\,(\mathbf{z},\ \mathbf{V}/(2\gamma_t)) \tag{11}$$

and show that $(\tau_t,\ \boldsymbol{\theta}_t^T,\ \mathbf{X}_{t-k+1,t})$ has a stationary distribution under which $(\boldsymbol{\theta}_t^T,\ \tau_t)$ has the same distribution as $(\mathbf{Z}_t^T,\ \gamma_t)$ in (11) and

$$X_t | (\boldsymbol{\theta}_t^T,\ \tau_t) \sim \text{Normal}\,(\mu_t/(1 - \rho_{1,t} - \cdots - \rho_{k,t}),\ \ (2\tau_t)^{-1} v_t)$$

where $v_t = \sum_{j=0}^{\infty} \beta_{j,t}^2$ and $\beta_{j,t}$ are the coefficients in the power series representation of $1/(1 - \alpha_{1,t} z - \cdots - \alpha_{k,t} z^k) = \sum_{j=0}^{\infty} \beta_{j,t} z^j$ for $|z| \le 1$. In addition, they show that the Markov chain $(\tau_t,\ \boldsymbol{\theta}_t^T,\ \mathbf{X}_{t-k+1,t})$ is reversible if it is initialized at the stationary distribution, and therefore the backward filter of $(\tau_t,\ \boldsymbol{\theta}_t^T)$ based on $X_n, \ldots, X_{t+1}$ has the same structure as the forward

predictor based on the past $n-t$ observations prior to t. Application of Bayes' theorem shows that the forward and backward filters can be combined to yield

$$f(\tau_t,\ \boldsymbol{\theta}_t|\mathbf{X}_{1,n})\ \propto f(\tau_t,\ \boldsymbol{\theta}_t|\mathbf{X}_{1,t})f(\tau_t,\ \boldsymbol{\theta}_t|\mathbf{X}_{t+1,n})/\pi(\tau_t,\ \boldsymbol{\theta}_t) \tag{12}$$

where π denotes the stationary density function, which is the same as that of $(\gamma_t,\ \mathbf{z}_t)$ given in (11); see Lai et al. (2005, p. 284).

Because the truncated normal is used in lieu of the normal distribution in (11), the conditional distribution of $\boldsymbol{\theta}_n$ given τ_n needs to be replaced by $\boldsymbol{\theta}_n|\tau_n \sim \mathbf{T}_C\ \text{Normal}\,(\mathbf{z}_{J_n,n}, \mathbf{V}_{J_n,n}/(2\tau_n))$, while (6) defining $\mathbf{V}_{j,n}$, $\mathbf{z}_{j,n}$ and $\alpha_{j,n}$ remains unchanged. For the implementation of the forward or backward filter, Lai et al. (2005) propose to ignore the truncation, as the constraint set C only serves to generate non-explosive observations but has little effect on the values of the weights $p_{j,n}$. Analogous to (5) and (6), the conditional distribution of $(\boldsymbol{\theta}_t, \tau_t)$ given $J_t = i$, $\tilde{J}_t = j+1$ and $\mathbf{X}_{i,j}$ for $i \le t < j \le n$ can be described by

$$\tau_t \sim \text{Gamma}\left(\text{g} + \frac{j-i+1}{2},\ \frac{1}{a_{i,j,t}}\right),\ \ \boldsymbol{\theta}_t|\tau_t \sim\ \ \text{Normal}\left(\text{Z}_{i,j,t,}\frac{1}{2\tau_t}\mathbf{V}_{i,j,t}\right) \tag{13}$$

if we ignore the truncation in the truncated normal distribution in (11), where

$$\begin{aligned}
\mathbf{V}_{i,j,t} &= (\mathbf{V}_{i,t}^{-1} + \tilde{\mathbf{V}}_{j,t}^{-1} - \mathbf{V}^{-1})^{-1}\\
\mathbf{z}_{i,j,t} &= \mathbf{V}_{i,j,t}(\mathbf{V}_{i,j,t}^{-1}\,\mathbf{z}_{i,t} + \tilde{\mathbf{V}}_{j,t}^{-1}\tilde{\mathbf{z}}_{j,t+1} - \mathbf{V}^{-1}\text{z})\\
\text{a}_{i,j,t} &= \lambda^{-1} + \mathbf{z}^T\mathbf{V}^{-1}\mathbf{z} + \sum_{l=i}^{j} X_l^2 - \mathbf{z}_{i,j,t}^T\mathbf{V}_{i,j,t}^{-1}\mathbf{z}_{i,j,t}
\end{aligned}$$

in which $\mathbf{V}_{i,t}$, $\mathbf{z}_{i,t}$ and $a_{i,t}$ are defined in (6) and $\tilde{\mathbf{V}}_{j,t}$, $\tilde{\mathbf{z}}_{j,t}$ and $\tilde{\text{a}}_{j,t}$ are defined similarly by reversing time. Let $|\cdot|$ denote the determinant of a matrix,

$$b_{i,j,t} = \left(\frac{|\mathbf{V}_{i,t}|\ |\tilde{\mathbf{V}}_{j,t}|}{|\mathbf{V}|\ |\mathbf{V}_{i,j,t}|}\right)^{-1/2}\left\{\frac{\Gamma(g)\Gamma(g+\frac{1}{2}(j-i+1))}{\Gamma(g+\frac{1}{2}(t-i+1))\Gamma(g+\frac{1}{2}(j-t))}\right\}\frac{a_{i,t}^{g+(t-i+1)/2}\tilde{a}_{j,t}^{g+(j-t)/2}}{a^g a_{i,j,t}^{g+(j-i+1)/2}}$$

$$B_t = p + (1-p)\sum_{k+1\le i\le t<j\le n} p_{i,t}\tilde{p}_{j,t}b_{i,\,j,t}$$

Using (12) and (13), Lai et al. (2005, p. 288) have shown that analogous to (7),

$$\begin{aligned} E(\sigma_t^2|\mathbf{X}_{1,n}) &\doteq \frac{p}{B_t}\sum_{i=k+1}^{t}\frac{p_{i,t}a_{i,t}}{2g+t-i-1}+\frac{1-p}{B_t}\sum_{k+1\le i\le t\le j\le n}p_{i,t}\tilde{p}_{i,t}b_{i,j,t}\frac{a_{i,j,t}}{2g+j-i+1} \\ E(\boldsymbol{\theta}_t|\mathbf{X}_{1,n}) &\doteq \frac{p}{B_t}\sum_{i=k+1}^{t}p_{i,t}\mathbf{z}_{i,t}+\frac{1-p}{B_t}\sum_{k+1\le i\le t\le j\le n}p_{i,t}\tilde{p}_{j,t}b_{i,j,t}\mathbf{z}_{i,j,t} \end{aligned} \tag{14}$$

in which the approximation ignores truncation within C. Moreover, the conditional probability of a structural break at time t ($\le n$) given $\mathbf{X}_{1,n}$ is

$$P\{I_t=1|\mathbf{X}_{1,n}\}=\sum_{i=k+1}^{t}P\{I_t=1,\ J_{t-1}=i|\mathbf{X}_{1,n}\}\doteq p/B_t \tag{15}$$

Although the Bayes filter uses a recursive updating formula (8) for the weights $p_{j,n}$ ($k<j\le n$), the number of weights increases with n, resulting in unbounded computational and memory requirements in estimating σ_n and $\boldsymbol{\theta}_n$ as n keeps increasing. Lai et al. (2005) propose a bounded complexity mixture (BCMIX) approximation to the optimal filter (7) by keeping the most recent m_p weights together with the largest n_p-m_p of the remaining weights at every stage n (which is tantamount to setting the other $n-n_p$ weights to be 0), where $0\le m_p<n_p$. For the forward and backward filters in the Bayes smoothers, we can again use these BCMIX approximations, yielding a BCMIX smoother that approximates (14) by allowing at most n_p weights $p_{i,t}$ and n_p weights $\tilde{p}_{j,t}$ to be nonzero. In particular, the Bayes estimates of the time-varying mean returns and volatilities in the bottom panels of Figs. 3 and 4 use $n_p=40$ and $m_p=25$.

3.3. Segmentation via Estimated Probabilities of Structural Breaks

The plot of the estimated probabilities of structural breaks over time in the bottom panel of Fig. 1 for the NASDAQ weekly returns provides a natural data segmentation procedure that locates possible change-points at times when these estimated probabilities are relatively high. These probabilities are computed via (15), in which the hyperparameters p, $\mathbf{z}$,$\mathbf{V}$,g and λ of the change-point autoregressive model (2)–(4) are determined as follows.

First, we fit an AR(2) model to $\{r_t, r_{t-1},\ldots,r_{t-19}\}$, $\max(k+1,\ 20)\le t\le n$, yielding $\tilde{\mu}_t,\tilde{\rho}_{1t},\tilde{\rho}_{2t}$ and $\tilde{\sigma}_t^2=t^{-1}\sum_{j=t-19}^{t}(r_j-\tilde{\mu}_t-\tilde{\rho}_{1t}r_{j-1}-\tilde{\rho}_{2t}r_{j-2})^2$, which are the "moving window estimates" (with window size 20) of μ_t, ρ_{1t}, ρ_{2t} and σ_t^2. The dotted curve in the bottom panel of Fig. 4 is a plot of the $\tilde{\sigma}_t$ series. Then we estimate $\mathbf{z}$ and $\mathbf{V}$ by the sample mean $\hat{\mathbf{z}}$ and the sample covariance

matrix $\hat{\mathbf{V}}$ of $\{(\tilde{\mu}_t, \tilde{\rho}_{1t}, \tilde{\rho}_{2t})^T, \max(k+1, 20) \le t \le n\}$, and apply the method of moments to estimate g and λ of (4). Specifically, regarding $(2\tilde{\sigma}_t^2)^{-1}$ as a sample from the Gamma(g, λ) distribution, $E(\tilde{\sigma}_t^2) = (2\lambda)^{-1}(g-1)^{-1}$ and $\mathrm{Var}(\tilde{\sigma}_t^2) = (2\lambda)^{-2}(g-1)^{-2}(g-2)^{-1}$, and using the sample mean and the sample variance of the $\tilde{\sigma}_t^2$ to replace their population counterparts yields the following estimates of g and λ:

$$\hat{g} = 2.4479, \quad \hat{\lambda} = 0.03960$$

$$\hat{\mathbf{z}} = (0.1304 \quad 0.003722 \quad -0.008111)^T,$$

$$\hat{\mathbf{V}} = \begin{pmatrix} 0.06096 & -0.009412 & 0.0003295 \\ -0.009412 & 0.01864 & 0.004521 \\ 0.0003295 & 0.004521 & 0.01023 \end{pmatrix}$$

With the hyperparameters $\mathbf{z}$, $\mathbf{V}$, g and λ thus chosen, we can estimate the hyperparameter p by maximum likelihood, noting that

$$f(X_t|\mathrm{X}_{1,t-1}) = pf(X_t|J_t = t) + (1-p)\sum_{j=k+1}^{t-1} P(J_{t-1} = j|\mathrm{X}_{1,t-1}) f(X_t|\mathrm{X}_{1,t-1}, J_{t-1} = j)$$

$$= \sum_{j=k+1}^{t} p_{j,t}^*$$

in which $p_{j,t}^*$ is defined by (8) and is a function of p. Hence the log-likelihood function is

$$l(p) = \log\left(\prod_{t=k+1}^{n} f(X_t|\mathrm{X}_{1,t-1})\right) = \sum_{t=k+1}^{n} \log\left(\sum_{j=k+1}^{t} p_{j,t}^*\right)$$

Figure 5 plots the log-likelihood function thus defined for the NASDAQ weekly returns from November 19, 1984 to September 15, 2003, giving the maximum likelihood estimate $\hat{p} = 0.027$ of the hyperparameter p.

4. NEAR UNIT ROOT BEHAVIOR OF LEAST SQUARES ESTIMATES IN A CHANGE-POINT AUTOREGRESSIVE MODEL

We have shown in Section 2 that volatility persistence in GARCH models may arise if the possiblity of structural changes is ignored. Similarly, spurious

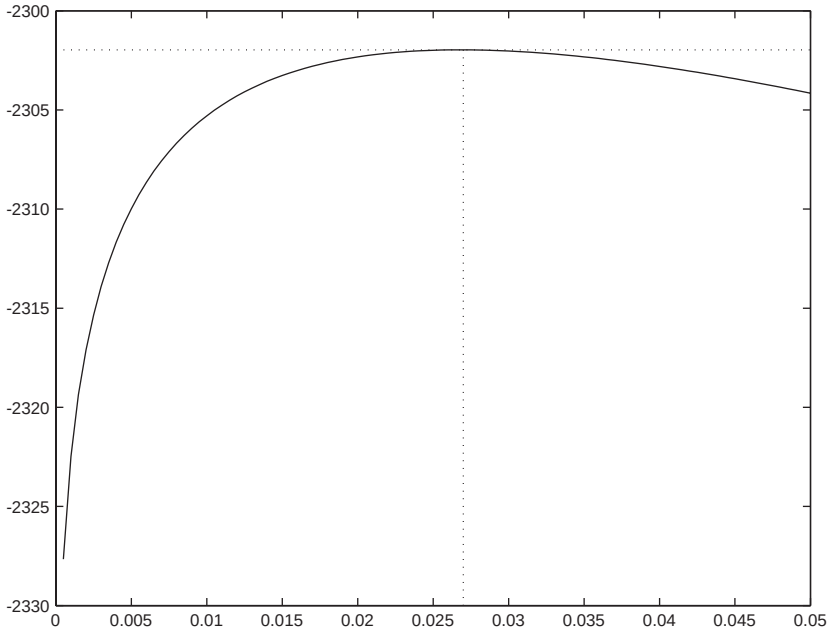


Fig. 5. Log-Likelihood Function of Hyperparameter p in Bayesian Change-Point Model.

unit root in autoregressive models may also arise if one omits the possibility of structural changes. There is an extensive literature on the interplay between unit root (or more general long memory) and structural breaks; see Perron (1989), Perron and Vogelsang (1992), Zivot and Andrews (1992), Lumsdaine and Papell (1997), Maddala, Kim, and Phillips (1999), Granger and Hyung (1999), Diebold and Inoue (2001) and the special issues of *Journal of Business and Economic Statistics* (1992) and *Journal of Econometrics* (1996).

In this Section, we use the HMM of structural changes similar to that in Section 3 to demonstrate that spurious long memory can arise when the possibility of structural change is not incorporated. Consider the autoregressive model with occasional mean shifts

$$X_n = \mu_n + \rho X_{n-1} + \varepsilon_n \tag{16}$$

where ε_n are i.i.d. unobservable standard normal random variables and are independent of $I_n := 1_{\{\mu_n \neq \mu_{n-1}\}}$, which are i.i.d. Bernoulli random variables

with $P(I_n = 1) = p$, as in (3), and

$$\mu_n = (1 - I_n)\mu_{n-1} + I_n Z_n \tag{17}$$

with i.i.d. normal Z_n that have common mean μ and variance v. The autoregressive parameter ρ is assumed to be less than 1 in absolute value but is unknown. The methods in Section 3.1 or 3.2 can be readily modified to estimate μ_n and ρ, either sequentially or from a given set of data $\{X_n, 1 \le n \le T\}$.

Without incorporating possible shifts in μ, suppose we fit an AR(1) model $X_n = \mu + \rho X_{n-1} + \varepsilon_n$ to $\{X_n, 1 \le n \le T\}$. The maximum likelihood estimate $(\hat{\mu}, \hat{\rho})$ is the same as the least squares estimate; in particular,

$$\hat{\rho} = \frac{\sum_{i=2}^{T} X_i X_{i-1} - \frac{1}{T-1}\left(\sum_{i=2}^{T} X_i\right)\left(\sum_{i=1}^{T-1} X_i\right)}{\sum_{i=1}^{T-1} X_i^2 - \frac{1}{T-1}\left(\sum_{i=1}^{T-1} X_i\right)^2} \tag{18}$$

Since $|\rho| < 1$, the Markov chain (μ_n, X_n) defined by (16) and (17) has a stationary distribution π under which μ_n has the $N(\mu, v)$ distribution and $(1 - \rho) X_n$ has the $N(\mu, 1+v)$ distribution. Hence

$$E_\pi(X_n) = \mu/(1 - \rho),\ E_\pi(X_n^2) = (1 + v + \mu^2)/(1 - \rho)^2,\ E_\pi(\mu_n) = \mu \tag{19}$$

Making use of (19) and the stationary distribution of (μ_n, X_n), we prove in the appendix the following result, which shows that the least squares $\hat{\rho}$ in the nominal model that ignores structural breaks can approach 1 as $n \to \infty$ even though $|\rho| < 1$.

Theorem. *Suppose that in* (16) and (17), $(1-p)\rho = 1 - pv$. *Then* $\hat{\rho} \to 1$ *with probability* 1.

Table 2 reports the results of a simulation study on the behavior of the least squares estimate $\hat{\rho}$ in fitting an AR(1) model with constant μ to $\{X_n, 1 \le n \le T\}$, in which X_n is generated from (16) that allows periodic jumps in μ. Unless stated otherwise, $T = 1000$. Each result in Table 2 is the average of 100 simulations from the change-point model (16), with the standard error included in parentheses. The table considers different values of the parameters p, ρ and v of the change-point model and also lists the corresponding values of the ratio $(1-p)\rho/(1-pv)$ in the preceding theorem. Table 2 shows that $\hat{\rho}$ is close to 1 when this ratio is 1, in agreement with the theorem. It also shows that $\hat{\rho}$ can be quite close to 1 even when this ratio differs substantially from 1. Also given in Table 2 are the p-values (computed by 'PP. test' in R) of the Phillips–Perron test of the null hypothesis that $\rho = 1$; see Perron (1988). These results show that near unit root

Table 2. Means and Standard Errors (in Parentheses) of Least Squares Estimates and P-Values of Unit Root Test.

p	v	ρ	$\frac{(1-p)\rho}{1-p^v}$	$\hat{\rho}$	P-Value	
0.002	4	0.05	0.5030	0.8628	0.1023	
				(0.0180)	(0.0174)	
0.002	8	0.20	0.2028	0.7591	0.1595	
				(0.0320)	(0.0240)	
0.002	8	0.84	0.8520	0.9730	0.4443	
				(0.0052)	(0.0371)	
0.002	16	0.50	0.5155	0.9134	0.3818	
				(0.0174)	(0.0329)	
0.004	8	0.50	0.5145	0.9505	0.3505	
				(0.0126)	(0.0318)	
0.008	8	0.50	0.5299	0.9891	0.3514	
				(0.0007)	(0.0252)	
0.01	16	84/99	1	0.9978	0.6110	
				(0.0002)	(0.0252)	
0.01	16	84/99	1	0.9986	0.4241	($T = 2000$)
				(0.0001)	(0.0201)	
0.01	16	84/99	1	0.9988	0.2585	($T = 3000$)
				(0.0001)	(0.0193)	
0.01	16	84/99	1	0.9988	0.0811	($T = 4000$)
				(0.0000)	(0.0090)	

behavior of $\hat{\rho}$ occurs quite frequently if the possibility of structural change is not taken into consideration.

5. CONCLUDING REMARKS

Financial time series modeling should incorporate the possibility of structural changes, especially when the series stretches over a long period of time. Volatility persistence or long memory behavior of the parameter estimates typically occurs if one ignores possible structural changes, as shown by the empirical study in Section 2 and the asymptotic theory in Section 4. Another way to look at such long memory behavior is through "multiresolution analysis" or "multiple time scales," in which the much slower time scale for parameter changes (causing "long memory") is overlaid on the short-run fluctuations of the time series.

The structural change model in Section 3 provides a powerful and tractable method to capture structural changes in both the volatility and regression parameters. It is a HMM, with the unknown regression and volatility

parameters $\boldsymbol{\theta}_t$ and σ_t undergoing Markovian jump dynamics, so that estimation of $\boldsymbol{\theta}_t$ and σ_t can be treated as filtering and smoothing problems in the HMM. Making use of the special structure of the HMM that involves gamma-normal conjugate priors, there are explicit recursive formulas for the optimal filters and smoothers. Besides the BCMIX approximation described in Section 3.2, another approach, introduced in Lai et al. (2005), is based on Monte Carlo simulations and involves sequential importance sampling with resampling. In Section 3.3, we have shown how the optimal structural change model can be applied to segment financial time series by making use of the estimated probabilities of structural breaks.

There is an extensive literature on tests for structural breaks in static regression models. Quandt (1960) and Kim and Siegmund (1989) considered likelihood ratio tests to detect a change-point in simple linear regression, and described approximations to the significance level. The issue of multiple change-points was addressed by Bai (1999), Bai and Perron (1998), and Bai, Lumsdaine, and Stock (1998). The computational complexity of the likelihood statistic and the associated significance level increase rapidly with the prescribed maximum number of change-points. Extension from static to dynamic regression models entails substantially greater complexity in the literature on tests for structural breaks in autoregressive models involving unit root and lagged variables; see Kramer, Plogerger, and Alt (1988), Banerjee, Lumsdaine, and Stock (1992), Doufour and Kivlet (1996), Harvey, Leybourne, and Newbold (2004) and the references therein. The posterior probabilities of change-points given in Section 3.2 for the Bayesian model (2)–(4), which incorporates changes not only in the autoregressive parameters but also in the error variances, provides a much more tractable Bayes test for structural breaks, allowing multiple change-points. The frequentest properties of the test are given in Lai and Xing (2005), where it is shown that the Bayes test provides a tractable approximation to the likelihood ratio test.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation.

REFERENCES

Albert, J. H., & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, *11*, 1–15.

Bai, J. (1999). Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, *91*, 299–323.

Bai, J., Lumsdaine, R. L., & Stock, J. H. (1998). Testing for and dating common breaks in multivariate time series. *Review of Economic Studies*, *65*, 395–432.

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, *66*, 47–78.

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

Banerjee, A., Lumsdaine, R. L., & Stock, J. H. (1992). Recursive and sequential tests of the unit-root and trend-break hypothesis: Theory and international evidence. *Journal of Business and Economic Statistics*, *10*, 271–287.

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

Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, *41*, 389–405.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*, 221–241.

Chib, S., Nardari, F., & Shepard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, *108*, 281–316.

Diebold, F. X. (1986). Comment on modeling the persistence of conditional variance. *Econometric Reviews*, *5*, 51–56.

Diebold, F. X., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, *105*, 131–159.

Doufour, J.-M., & Kivlet, J. F. (1996). Exact tests for structural change in first-order dynamic models. *Journal of Econometrics*, *70*, 39–68.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*, 987–1007.

Engle, R. F., & Bollerslev, T. (1986). Modeling the persistence of conditional variances. *Econometric Reviews*, *5*, 1–50.

Granger, C. W. J., & Hyung, N. (1999).*Occasional structural breaks and long memory*. Economics working paper series, 99-14, Department of Economics, UC San Diego.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*, 357–384.

Hamilton, J. D., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*, 307–333.

Harvey, D. L., Leybourne, S. J., & Newbold, P. (2004). Tests for a break in level when the order of integration is unknown. *Oxford Bulletin of Economics and Statistcs*, *66*, 133–146.

Kim, H., & Siegmund, O. D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, *76*, 409–423.

Kramer, W., Plogerger, W., & Alt, R. (1988). Testing for structural change in dynamic models. *Econometrica*, *56*, 1355–1369.

Lai, T. L., Liu, H., & Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica*, *15*, 279–301.

Lai, T. L., & Xing, H. (2005). *Hidden Markov models for structural changes and stochastic cycles in regression and time series*. Working paper, Department of Statistics, Stanford University.

Lamoureux, C. G., & Lastraps, W. D. (1990). Persistence in variance, structural change and the GARCH model. *Journal of Business and Economic Statistics*, *8*, 225–234.

Lumsdaine, R. L., & Papell, D. H. (1997). Multiple trend breaks and the unit root hypothesis. *Review of Economics and Statistics*, *79*, 212–218.

Maddala, G. S., Ki, I., & Phillips, P. C. B. (1999). *Unit roots, cointegration, and structural change*. New York: Cambridge University Press.

McCulloch, R. E., & Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, *88*, 968–978.

Meyn, S. P., & Tweedie, R. L. (1993). *Markov chains and stochastic stability*. New York: Springer-Verlag.

Perron, P. (1988). Trends and random walks in macroeconomic time series. *Journal of Economic Dynamics and Control*, *12*, 297–332.

Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, *57*, 1361–1401.

Perron, P., & Vogelsang, T. J. (1992). Nonstationarity and level shifts with an application to purchasing power parity. *Journal of Business and Economic Statistics*, *10*, 301–320.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two seperate regimes. *Journal of the American Statistical Association*, *55*, 324–330.

Wang, J., & Zivot, E. (2000). A Bayesian time series model of multiple structural changes in level, trend and variance. *Journal of Business and Economic Statistics*, *18*, 374–386.

Zivot, E., & Andrews, D. W. K. (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics*, *10*, 251–270.

APPENDIX

The proof of the theorem makes use of (18), (19) and

$$\mathrm{Cov}_\pi(\mu_{n+1}, X_n) = (1-p)v/\{1-(1-p)\rho\} \tag{A.1}$$

To prove (A.1), first note from (16) that

$$\begin{aligned}\mathrm{Cov}_\pi(\mu_n, X_n) &= v + \rho\mathrm{Cov}(\mu_n, X_{n-1}) \\ &= v + \rho\{E(\mu_n, X_{n-1}) - E(\mu_n)E(X_{n-1})\} \\ &= v + (1-p)\rho\mathrm{Cov}_\pi(\mu_{n-1}, X_{n-1})\end{aligned} \tag{A.2}$$

The last equality in (A.2) follows from (17) and that $E_\pi(I_n Z_n) = p\mu$. The recursive representation of $\mathrm{Cov}_\pi(\mu_n, X_n)$ in (A.2) yields

$$\mathrm{Cov}_\pi(\mu_n, X_n) = v/\{1-(1-p)\rho\}$$

Combining this with the first equality in (A.2) establishes (A.1). In view of (A.2) and the strong law for Markov random walks (cf. Theorem 17.1.7 of

Meyn and Tweedie (1993)),

$$\sum_{i=1}^{T-1} X_i/T \to E_\pi(X_n),$$

$$\left(\sum_{i=1}^{T-1} X_{i+1}X_i\right)/T = \left\{\sum_{i=1}^{T-1} \mu_{i+1}X_i + \rho\sum_{i=1}^{T-1} X_i^2 + \sum_{i=1}^{T-1} \varepsilon_{i+1}X_i\right\}/T$$

$$\to E_\pi(\mu_{n+1}X_n) + \rho E_\pi(X_n^2) = \mathrm{Cov}_\pi(\mu_{n+1}, X_n)$$

$$+ E_\pi(\mu_{n+1})E_\pi(X_n) + \rho E_\pi(X_n^2) \tag{A.3}$$

with probability 1, noting that $E_\pi(X_n\varepsilon_{n+1}) = 0$. Putting (A.3) and (19), (A.1) into (18) shows that with probability 1, $\hat{\rho}$ converges $\rho + \{v(1-p)(1-p)^2\}/\{(1+v)[1-(1-p)\rho]\}$, which is equal to 1 if $\rho = (1-pv)/(1-p)$.

TIME SERIES MEAN LEVEL AND STOCHASTIC VOLATILITY MODELING BY SMOOTH TRANSITION AUTOREGRESSIONS: A BAYESIAN APPROACH

Hedibert Freitas Lopes and Esther Salazar

ABSTRACT

In this paper, we propose a Bayesian approach to model the level and the variance of (financial) time series by the special class of nonlinear time series models known as the logistic smooth transition autoregressive models, or simply the LSTAR models. We first propose a Markov Chain Monte Carlo (MCMC) algorithm for the levels of the time series and then adapt it to model the stochastic volatilities. The LSTAR order is selected by three information criteria: the well-known AIC and BIC, and by the deviance information criteria, or DIC. We apply our algorithm to a synthetic data and two real time series, namely the canadian lynx data and the SP500 returns.

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 225–238
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20028-2

1. INTRODUCTION

Smooth transition autoregressions (STAR), initially proposed in its univariate form by Chan and Tong (1986) and further developed by Luukkonen, Saikkonen, and Teräsvirta (1988) and Teräsvirta (1994), have been extensively studied over the last 20 years in association to the measurement, testing, and forecasting of nonlinear financial time series. The STAR model can be seen as a continuous mixture of two AR(k) models, with the weighting function defining the degree of nonlinearity. In this article, we focus on an important subclass of the STAR model, the logistic STAR model of order k, or simply the logistic smooth transition auto regressive LSTAR(k) model, where the weighting function has the form of a logistic function.

Dijk, Teräsvirta, and Franses (2002) make an extensive review of recent developments related to the STAR model and its variants. From the Bayesian point of view, Lubrano (2000) used weighted resampling schemes (Smith & Gelfand, 1992) to perform exact posterior inference of both linear and nonlinear parameters. Our main contribution is to propose an Markov Chain Monte Carlo (MCMC) algorithm that allows easy and practical extensions of LSTAR models to modeling univariate and multivariate stochastic volatilities. Therefore, we devote Section 2 to introducing the general LSTAR(k) to model the levels of a time series. Prior specification, posterior inference, and model choice are also discussed in this section. Section 3 adapts the MCMC algorithm presented in the previous section to modeling stochastic volatility. Finally, in Section 4, we apply our estimation procedures to a synthetic data and two real time series, namely the canadian lynx data and the SP500 returns, with final thoughts and perspectives listed in Section 5.

2. LSTAR: MEAN LEVEL

Let y_t be the observed value of a time series at time t and $x_t = (1, y_{t-1}, \ldots, y_{t-k})'$ the vector of regressors corresponding to the intercept plus k lagged values, for $t = 1, \ldots, n$. The logistic smooth transition autoregressive model of order k, or simply LSTAR(k), is defined as follows:

$$y_t = x_t'\theta_1 + F(\gamma, c, s_t)x_t'\theta_2 + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma^2) \tag{1}$$

with F playing the role of a smooth transition continuous function bounded between 0 and 1. In this paper, we focus on the logistic transition, i.e.

$$F(\gamma, c, s_t) = \{1 + \exp(-\gamma(s_t - c))\}^{-1}. \tag{2}$$

Several other functions could be easily accommodated, such as the exponential or the second order logistic function (Dijk et al., 2002). The parameter $\gamma > 0$ is responsible for the smoothness of F, while c is a location or *threshold* parameter. When $\gamma \to \infty$, the LSTAR model reduces to the well-known self-exciting TAR (SETAR) model (Tong, 1990) and when $\gamma = 0$ the standard AR(k) model arises. Finally, s_t is called the transition variable, with $s_t = y_{t-d}$ commonly used (Teräsvirta, 1994), and d a delay parameter. Even though we use y_{t-d} as the transition variable throughout this paper, it is worth emphasizing that any other linear/nonlinear functions of exogenous/endogenous variables can be easily included with minor changes in the algorithms presented and studied in this paper. We will also assume throughout the paper, without loss of generality, that $d \leq k$ and $y_{-k+1}, \ldots, y_0$ are known and fixed quantities. This is common practice in the time series literature and would only marginally increase the complexity of our computation and could easily be done by introducing a prior distribution for those missing values.

The LSTAR(k) model can be seen as model that allows smooth and continuous shifts between two extreme regimes. More specifically, if $\delta = \theta_1 + \theta_2$, then $x_t'\theta_1$ and $x_t'\delta$ represent the conditional means under the two extreme regimes

$$y_t = (1 - F(\gamma, c, y_{t-d}))x_t'\theta_1 + F(\gamma, c, y_{t-d})x_t'\delta + \varepsilon_t \tag{3}$$

or

$$y_t = \begin{cases} x_t'\theta_1 + \varepsilon_t & \text{if } \omega = 0 \\ (1-\omega)x_t'\theta_1 + \omega x_t'\delta + \varepsilon_t & \text{if } 0 < \omega < 1 \\ x_t'\delta + \varepsilon_t & \text{if } \omega = 1 \end{cases}$$

Therefore, the model (3) can be expressed as (1) with $\theta_2 = \delta - \theta_1$ being the contribution to the regression of considering a second regime. Huerta, Jiang, and Tanner (2003) also address issues about nonlinear time series models through logistic functions. We consider the following parameterization that will be very useful in the computation of the posterior distributions:

$$y_t = z_t'\theta + \varepsilon_t \tag{4}$$

where $\theta' = (\theta_1', \theta_2')$ represent the linear parameters, (γ, c) are the nonlinear parameters, $\Theta = (\theta, \gamma, c, \sigma^2)$ and $z_t' = (x_t', F(\gamma, c, y_{t-d})x_t')$. The dependence of z_t on γ,c and y_{t-d} will be made explicit whenever necessary. The likelihood is

then written as

$$p(\boldsymbol{y}|\Theta) \propto \sigma^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - z_t'\theta)^2\right\} \tag{5}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$.

2.1. Prior Specification and Posterior Inference

In this Section, we derive MCMC algorithm for posterior assessment of both linear and nonlinear parameters and for the observations' variance, σ^2, so that the number of parameters to be estimated is $2k+5$. We also consider the estimation of the delay parameter d. We adopt Lubrano's (2000) formulation, i.e. $(\theta_2|\sigma^2, \gamma) \sim N(0, \sigma^2 e^{\gamma} I_{k+1})$ and $\pi(\theta_1, \gamma, \sigma^2, c) \propto (1+\gamma^2)^{-1}\sigma^{-2}$ for $\theta_1 \in \Re^{k+1}, \gamma, \sigma^2 > 0$ and $c \in [c_a, c_b]$, such that the conditional prior for θ_2 becomes more informative as γ approaches zero, relatively noninformative prior about θ_1 and γ, noninformative prior about σ^2, and relatively noninformative prior about c, with $c_a = \hat{\mathscr{F}}^{-1}(0.15)$, $c_b = \hat{\mathscr{F}}^{-1}(0.85)$ and $\hat{\mathscr{F}}$ the data's empirical cumulative distribution function. The prior density is then

$$\pi(\Theta) \propto \sigma^{-3}(1+\gamma^2)^{-1} \exp\left\{-\frac{1}{2}[\gamma + \sigma^{-2}e^{-r}\theta_2'\theta_2]\right\} \tag{6}$$

which combined with the likelihood (Eq. (5)) yields the posterior distribution

$$\pi(\Theta|y) \propto \frac{\sigma^{-(n+6)/2}}{(1+\gamma^2)} \exp\left\{-\frac{1}{2\sigma^2}[\gamma\sigma^2 + e^{-\gamma}\theta_2'\theta_2 + \sum_{t=1}^{n}(y_t - z_t'\theta)^2]\right\} \tag{7}$$

with Θ *in* $A = \left\{\Re^{2k+2} \times \Re^{+} \times \Re \times \Re^{+}\right\}$.

Needless to mention that the posterior distribution has no known form and that inference is facilitated by MCMC methods (Gilks, Richardson, & Spiegelhalter, 1996). The full conditional posterior of θ and σ^2 are respectively, normal and inverse gamma, in which case the application of Gibbs steps are straigthforward (Gelfand & Smith, 1990). However, the full conditional distributions of γ and c are of unknown form, so we use Metropolis–Hastings steps (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Hastings, 1970). For more details about Gibbs and Metropolis–Hastings algorithms see Gilks et al. (1996) and the books by Gamerman (1997) and Robert and Casella (1999).

2.2. Choosing the Model Order k

In this Section, we introduce the traditional information criteria, Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978), widely used in model selection/comparison. To this toolbox we add the DIC (Deviance Information Criterion) that is a criterion recently developed and widely discussed in Spiegelhalter, Best, Carlin, and Linde (2002).

Traditionally, model order or model specification are compared through the computation of information criteria, which are well known for penalizing likelihood functions of overparametrized models. AIC (Akaike, 1974) and BIC (Schwarz, 1978) are the most used ones. For data y and parameter θ, these criteria are defined as follows: $AIC = -2\ln(p(\boldsymbol{y}|\hat{\theta})) + 2p$ and $BIC = -2\ln(p(\boldsymbol{y}|\hat{\theta})) + p\ \ln\ n$, p is the dimension of θ, with $p = 2k + 5$ for LSTAR(k), sample size n and maximum likelihood estimator, $\hat{\theta}$. One of the major problems with AIC/BIC is that to define k is not trivial, mainly in Bayesian hierarchical models, where the priors act like reducers of the effective number of parameters through its interdependencies. To overcome this limitation, Spiegelhalter et al. (2002) developed an information criterion that properly defines the effective number parameter by $p_D = \bar{D} - D(\tilde{\theta})$, where $D(\theta) = -2\ln p(\boldsymbol{y}|\theta)$ is the deviance, $\tilde{\theta} = E(\theta|\boldsymbol{y}|)$ and $\bar{D} = E(D(\theta)|y)$. As a by-product, they proposed the *Deviance Information Criterion*: $DIC = D(\tilde{\theta}) + 2p_D = \bar{D} + p_D$. One could argue that the most attractive and appealing feature of the DIC is that it combines model fit (measured by $\bar{D}$) with model complexity (measured by p_D). Besides, DICs are more attractive than Bayes Factors since the former can be easily incorporated into MCMC routines. For successful implementations of DIC we refer to Zhu and Carlin (2000) (spatio-temporal hierarchical models) and Berg, Meyer, and Yu (2004) (stochastic volatility models), to name a few.

3. LSTAR: STOCHASTIC VOLATILITY

In this Section, we adopt the LSTAR structure introduced in Section 2 to model time-varying variances, or simply the *stochastic volatility*, of (financial) time series. It has become standard to assume that y_t, the observed value of a (financial) time series at time t, for $t = 1, \ldots, n$, is normally distributed conditional on the unobservable volatility h_t, i.e. $y_t|h_t \sim N(0, e^{h_t})$, and that the log-volatilities, $\lambda_t = \log h_t$, follow an autoregressive process of order one, i.e. $\lambda_t \sim N(\theta_0 + \theta_1\lambda_{t-1}; \sigma^2)$, with θ_0, θ_1 and σ^2

interpreted as the volatility's level, persistence and volatility, respectively. Several variations and generalizations based on this AR(1) structure were proposed over the last two decades to accommodate asymmetry, heavy tails, strong persistence and other features claimed to be presented in the volatility of financial time series (Kim, Shephard, & Chib, 1998).

As mentioned earlier, our contribution in this section is to allow the stochastic volatility to evolve according to the LSTAR(k) model, Eq. (1) from Section 2

$$\lambda_t = x_t'\theta_1 + F(\gamma, c, \lambda_{t-d})x_t'\theta_2 + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma^2) \tag{8}$$

with $x_t' = (1, \lambda_{t-1}, \ldots, \lambda_{t-k})$, θ_1 and $\theta_2(k+1)$-dimensional vectors and $F(\gamma, c, \lambda_{t-d})$ the logistic function introduced in Eq. (2). We assume, without loss of generality, that d is always less than or equal to k. In the particular case, when $k = 1$, the AR(1) model is replaced by an LSTAR(1)

$$\lambda_t \sim N\{\theta_{10} + \theta_{11}\lambda_{t-1} + [\theta_{20} + \theta_{21}\lambda_{t-1}]F(\gamma, c, \lambda_{t-d}); \sigma^2\}$$

Conditional on the vector of log-volatilities, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, we adopt, for the static parameters describing the LSTAR model, the same prior structure presented in Section 2.1. In other words, $(\theta_2|\sigma^2, \gamma) \sim N(0, \sigma^2 e^{\gamma} I_{K+1})$ and $\pi(\theta_1, \gamma, \sigma^2, c) \propto (1+\gamma^2)^{-1}\sigma^{-2}$ for $\theta_1 \in \Re^{k+1}, \gamma, \sigma^2 > 0$ and $c \in [\hat{\mathscr{F}}^{-1}(0.15), \hat{\mathscr{F}}^{-1}(0.85)]$, with $\hat{\mathscr{F}}$ an approximation to the log-volatilities' empirical cdf. Therefore, conditional on $\boldsymbol{\lambda}$, sampling θ_1, θ_2, c, γ and σ^2 follows directly from the derivations of Section 2.1. Posterior inference to this stochastic volatility LSTAR(k), or simply SV-LSTAR(k), follows closely the developments of Section 2.1 and the Appendix, with the additional burden of sampling the hidden vector of log-stochastic volatilities, $\boldsymbol{\lambda}$. We use the single parameter move introduced by Jacquier, Polson, and Rossi (1994). Specifically, we rely on the fact the full conditional distribution of λ_t given the parameters in Θ plus $\boldsymbol{\lambda}_{-t} = \{\lambda_1, \ldots, \lambda_{t-1}, \lambda_{t+1}, \ldots, \lambda_n\}$ is given by

$$p(\lambda_t|\boldsymbol{\lambda}_{-t}, \Theta, \boldsymbol{y}) \propto g(\lambda_t|\boldsymbol{\lambda}_{-t}, \Theta, \boldsymbol{y}) \equiv p(y_t|\lambda_t)\prod_{i=0}^{k} p(\lambda_{t+i}|\lambda_{t+i-1}, \ldots, \lambda_{t+i-k,}, \Theta) \tag{9}$$

which has no known closed form. We sample λ_t^* from $N(\lambda_t^{(j)}, \Delta)$, for Δ a problem-specific tuning parameter, and $\lambda_t^{(j)}$ the current value of λ_t in this Random-Walk Markov chain. The draw λ_t^* is accepted with probability.

$$\alpha = \min\left\{1, \frac{g(\lambda_t^*|\boldsymbol{\lambda}_{-t,}, \Theta, \boldsymbol{y})}{g(\lambda_t^{(j)}|\boldsymbol{\lambda}_{-t}, \Theta, \boldsymbol{y})}\right\}$$

We do not claim that our proposal is optimal or efficient by any formal means, but we argue that they are useful from a practical viewpoint as our simulated and real data applications reveal.

4. APPLICATIONS

In this Section, our methodology is extensively studied against simulated and real time series. We start with an extensive simulation study, which is followed by the analysis of two well known dataset: (i) The *Canadian Lynx* series, which stands for the number of Canadian Lynx trapped in the Mackenzie River district of North-west Canada, and (ii) The USPI Index, which stands for the US Industrial Production Index.

4.1. A Simulation Study

We performed a study by simulating a time series with 1,000 observations from the following the LSTAR(2):

$$y_t = 1.8y_{t-1} - 1.06y_{t-2} + (0.02 - 0.9y_{t-1} + 0.795y_{t-2})F(100, 0.02, y_{t-2}) + \varepsilon_t \tag{10}$$

where $F(100, 0.02, y_{t-2}) = [1 + \exp\{-100(y_{t-2} - 0.02)\}]^{-1}$ and $\varepsilon_t \sim N(0, 0.02^2)$. The initial values were $k = 5$, $\theta_1 = \theta_2 = (0, 1, 1, 1, 1, 1)$, $\gamma = 150$, $c = \bar{y}$ and $\sigma^2 = 0.01^2$. We consider 5,000 MCMC runs with the first half used as burn-in. AIC, BIC and DIC statistics are presented in Table 1.

Posterior means and standard deviations for the model with highest posterior model probability are shown in Table 2. The three information criteria point to the correct model, while our MCMC algorithm produces fairly accurate estimates of posterior quantities.

4.2. Canadian Lynx

We analyze the well-known *Canadian Lynx* series. Figure 1 shows the logarithm of the number of Canadian Lynx trapped in the Mackenzie River district of North-west Canada over the period from 1821 to 1934 (data can be found in Tong, 1990, p. 470). For further details and previous analysis of this time series, see Ozaki (1982), Tong (1990), Teräsvirta (1994), Medeiros and Veiga (2005), and Xia and Li (1999), among others.

Table 1. Model Comparison using Information Criteria for the Simulated LSTAR(2) Process.

k	d	*AIC*	*BIC*	*DIC*
1	1	−4654.6	−4620.2	−8622.2
	2	−4728.6	−4694.3	−8731.2
	3	−4682.7	−4648.4	−8606.8
2	1	−3912.0	−3867.9	−7434.2
	2	−5038.9	−4994.8	−9023.6
	3	−4761.2	−4717.0	−8643.8
3	2	−5037.0	−4983.0	−9023.3
	3	−4850.4	−4796.5	−8645.9

Table 2. Posterior Means for the Parameters by LSTAR(2) Model using 2,500 MCMC Runs.

Par	True Value	Mean	St.Dev.	Par	True Value	Mean	St.Dev.
θ_{01}	0	−0.0028	0.0021	θ_{02}	0.02	0.023	0.0036
θ_{11}	1.8	1.7932	0.0525	θ_{12}	−0.9	−0.8735	0.0637
θ_{21}	−1.06	−1.0809	0.0654	θ_{22}	0.795	0.7861	0.0746
γ	100	100.87	4.9407	c	0.02	0.0169	0.0034
σ^2	0.0004	0.00037	0.000016				

We run our MCMC algorithm for 50,000 iterations and discard the first half as burn-in. The LSTAR(11) model with $d = 3$ has the highest posterior model probability when using the BIC as a proxy to Bayes factor

$$
\begin{aligned}
y_t = {} & \overset{(0.307)}{0.987} + \overset{(0.111)}{0.974}\, y_{t-1} - \overset{(0.151)}{0.098}\, y_{t-2} - \overset{(0.142)}{0.051}\, y_{t-3} - \overset{(0.137)}{0.155}\, y_{t-4} + \overset{(0.143)}{0.045}\, y_{t-5} \\
& - \overset{(0.146)}{0.0702}\, y_{t-6} - \overset{(0.158)}{0.036}\, y_{t-7} + \overset{(0.167)}{0.179}\, y_{t-8} + \overset{(0.159)}{0.025}\, y_{t-9} + \overset{(0.144)}{0.138}\, y_{t-10} - \overset{(0.096)}{0.288}\, y_{t-11} \\
& + (\overset{(2.04)}{-3.688} + \overset{(0.431)}{1.36}\, y_{t-1} - \overset{(0.744)}{3.05}\, y_{t-2} + \overset{(1.111)}{4.01}\, y_{t-3} - \overset{(0.972)}{2.001}\, y_{t-4} + \overset{(0.753)}{1.481}\, y_{t-5} \\
& + \overset{(0.657)}{0.406}\, y_{t-6} - \overset{(0.735)}{0.862}\, y_{t-7} - \overset{(0.684)}{0.666}\, y_{t-8} + \overset{(0.539)}{0.263}\, y_{t-9} + \overset{(0.486)}{0.537}\, y_{t-10} - \overset{(0.381)}{0.569}\, y_{t-11} \\
& \times \left(1 + \exp\{-\overset{(0.688)}{11.625}(y_{t-8} - \overset{(0.017)}{3.504})\}\right)^{-1} + \varepsilon_t, \qquad E(\sigma^2|y) = 0.025
\end{aligned}
$$

with posterior standard deviations in parentheses. The LSTAR($k = 11$) captures the decennial seasonality exhibited by the data. The vertical lines on Part (a) of Fig. 1 are estimates of the logistic transition function, $F(\gamma, c, y_{t-3})$

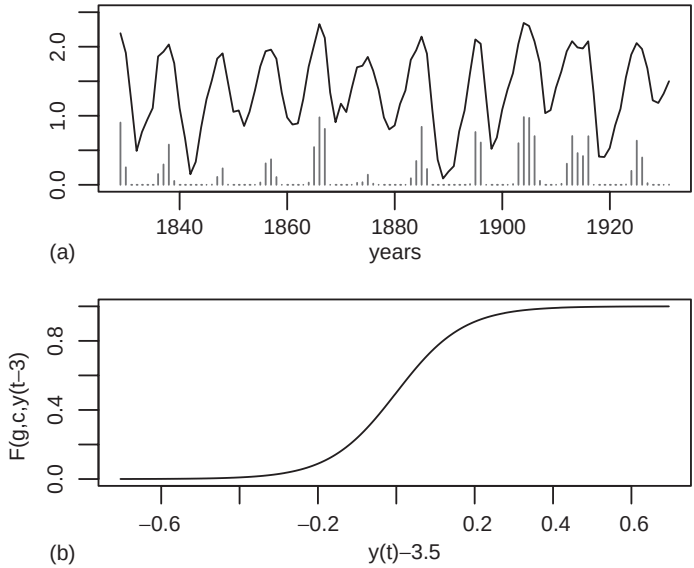


Fig. 1. Canadian Lynx Series: (a) Logarithm of the Number of Canadian Lynx trapped in the Mackenzie River District of North-West Canada over the Period from 1821 to 1934 (Time Series is Shifted 1.5 Units down to Accommodate the Transition Function). The Vertical Lines are Estimates of the Logistic Transition Function, $F(\gamma, c, y_{t-3})$ from Eq. (2), for $\gamma = -11.625$ and $c = 3.504$. Posterior Means of γ and c, respectively, (b) The Posterior Mean of the Transition Function as a Function of $y-3.5$.

from Eq. 2, for $\gamma = -11.625$ and $c = 3.504$ posterior means of γ and c, respectively. Roughly speaking, within each decade the transition cycles according to both extreme regimes as y_t gets away from $c = 3.504$, or the number of trapped lynx gets away from 3,191, which happens around 15% of the time. Part (b) of Fig. 1 exhibits the transition function as a function of $y-3.5$, which is relatively symmetric around zero corroborating with the fact that the decennial cycles are symmetric as well. Similar results were found in Medeiros and Veiga (2005) who fit an LSTAR(2) with $d = 2$ and Teräsvirta (1994) who fit an LSTAR(11) with $d = 3$ for the same time series.

4.3. S&P500 Index

Here, we fit the LSTAR(k)-stochastic volatility model proposed in Section 3 to the North American Standard and Poors 500 index, or simply the SP500 index, which was observed from January 7th, 1986 to December 31st, 1997,

a total of 3,127 observations. We entertained six models for the stochastic volatility and did model comparison by using the three information criteria presented in Section 2.2, i.e., AIC, BIC and DIC. The six models are: $\mathscr{M}_1$: $AR(1)$,$\mathscr{M}_2$: $AR(2)$,$\mathscr{M}_3$: $LSTAR(1)$ with $d = 1$, $\mathscr{M}_4$: $LSTAR(1)$ with $d = 2$, $\mathscr{M}_5$: $LSTAR(2)$ with $d = 1$, and $\mathscr{M}_6$: $LSTAR(2)$ with $d = 2$.

Table 3 presents the performance of the six models, in which all three information criteria agree upon the choice of the *best* model for the log-volatilities: $LSTAR(1)$ with $d = 1$. One can argue that the linear relationship prescribed by an AR(1) structure is insufficient to capture the dynamic behavior of the log-volatilities. The LSTAR structure brings more flexibility to the modeling. Table 4 presents the posterior mean and standard devations of all parameters for each one of the six models listed above.

5. CONCLUSIONS

In this paper, we developed Markov Chain Monte Carlo (MCMC) algorithms for posterior inference in a broad class of nonlinear time series models known as logistic smooth transition autoregressions, LSTAR. Our developments are checked against simulated and real dataset with encouraging results when modeling both the levels and the variance of univariate time series with LSTAR structures. We used standard information criteria such as the AIC and BIC to compare models and introduced a fairly new Bayesian criterion, the Deviance Information Criteria (DIC), all of which point to the same direction in the examples we explored in this paper.

Even though we concentrated our computations and examples to the logistic transition function, the algorithms we developed can be easily adapted to other functions such as the exponential or the second-order

Table 3. S&P500: Model Comparison based on Information Criteria: AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and DIC (Deviance Information Criterion).

Models	AIC	BIC	DIC
$\mathscr{M}_1$: $AR(1)$	12795	31697	7223.1
$\mathscr{M}_2$: $AR(2)$	12624	31532	7149.2
$\mathscr{M}_3$: $LSTAR(1, d = 1)$	12240	31165	7101.1
$\mathscr{M}_4$: $LSTAR(1, d = 2)$	12244	31170	7150.3
$\mathscr{M}_5$: $LSTAR(2, d = 1)$	12569	31507	7102.4
$\mathscr{M}_6$: $LSTAR(2, d = 2)$	12732	31670	7159.4

Table 4. S&P500: Posterior Means and Posterior Standard Deviations for the Parameters from all Six Entertained Models: $\mathscr{M}_1$: AR(1), $\mathscr{M}_2$: AR(2), $\mathscr{M}_3$: LSTAR(1) With d = 1, $\mathscr{M}_4$: LSTAR(1) With d = 2, $\mathscr{M}_5$: LSTAR(2) With d = 1, and $\mathscr{M}_6$: LSTAR(2) with d = 2.

Parameter	Models					
	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$	$\mathscr{M}_4$	$\mathscr{M}_5$	$\mathscr{M}_6$
	Posterior mean (standard deviation)					
θ_{01}	−0.060 (0.184)	−0.066 (0.241)	0.292 (0.579)	−0.354 (0.126)	−4.842 (0.802)	−6.081 (1.282)
θ_{11}	0.904 (0.185)	0.184 (0.242)	0.306 (0.263)	0.572 (0.135)	−0.713 (0.306)	−0.940 (0.699)
θ_{21}	—	0.715 (0.248)	—	—	−1.018 (0.118)	−1.099 (0.336)
θ_{02}	—	—	−0.685 (0.593)	0.133 (0.092)	4.783 (0.801)	6.036 (1.283)
θ_{12}	—	—	0.794 (0.257)	0.237 (0.086)	0.913 (0.314)	1.091 (0.706)
θ_{22}	—	—	—	—	1.748 (0.114)	1.892 (0.356)
γ	—	—	118.18 (16.924)	163.54 (23.912)	132.60 (10.147)	189.51 (0.000)
c	—	—	−1.589 (0.022)	0.022 (0.280)	−2.060 (0.046)	−2.125 (0.000)
σ^2	0.135 (0.020)	0.234 (0.044)	0.316 (0.066)	0.552 (0.218)	0.214 (0.035)	0.166 (0.026)

logistic functions, or even combinations of those. Similarly, even though we chose to work with y_{t-d} as the transition variable, our findings are naturally extended to situations where y_{t-d} is replaced by, say, $s(y_{t-1}, \ldots, y_{t-d}, \alpha)$ for α and d unknown quantities.

In this paper, we focused on modeling the level and the variance of univariate time series. Our current research agenda includes (i) the generalization of the methods proposed here to model factor stochastic volatility problems (Lopes & Migon, 2002; Lopes, Aguilar, and West, 2000), (ii) fully treatment of k as another parameter and posterior inference based on reversible jump Markov Chain Monte Carlo (RJMCMC) algorithms for both the levels and the stochastic volatility of (financial) time series (Lopes

& Salazar, 2005), and (iii) comparing smooth transition regressions with alternative jump models, such as Markov switching models (Carvalho & Lopes, 2002).

All our computer codes are available upon request and are fully programmed in the student's version of Ox, a statistical language that can be downloaded free of charge from http://www.nuff.ox.ac.uk/Users/Doornik/index.html.

ACKNOWLEDGMENTS

Hedibert Lopes would like to thank the Graduate School of Business, University of Chicago and the Department of Statistical Methods, Federal University of Rio de Janeiro for supporting his research on this paper. Esther Salazar was partially supported by a scholarship from CNPq, *Conselho Nacional de Desenvolvimento Científico e Tecnológico*, to pursue her master's degree in Statistics at the Federal University of Rio de Janeiro. We thank the comments of the participants of III Annual Advances in Econometrics Conference (USA), Science of Modeling and Symposium on Recent Developments in Latent Variables Modelling (Japan), II Congreso Bayesiano de America Latina (Mexico) and VII Brazilian Meeting of Bayesian Statistics, and the participants of talks delivered at the Catholic University of Rio de Janeiro and Federal University of Paraná (Brazil), and Northern Illinois University (USA). We also would like to thank the invaluable comments made by the referee.

REFERENCES

Akaike, H. (1974). New look at the statistical model identification. *IEEE Transactions in Automatic Control AC*, *19*, 716–723.

Berg, A., Meyer, R., & Yu, J. (2004). DIC as a model comparison criterion for stochastic volatility models. *Journal of Business & Economic Statistics*, *22*, 107–120.

Carvalho, C., & Lopes, H. F. (2002). *Simulation-based sequential analysis of Markov switching stochastic volatility models*. Technical Report. Department of Statistical Methods, Federal University of Rio de Janeiro.

Chan, K. S., & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, *7*, 179–190.

van Dijk, D., Teräsvirta, T., & Franses, P. (2002). Smooth transition autoregressive models – a survey of recent developments. *Econometrics Reviews*, *21*, 1–47.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. London: Chapman & Hall.

Gelfand, A. E., & Smith, A. M. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Hastings, W. R. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Huerta, G., Jiang, W., & Tanner, M. (2003). Time series modeling via hierarchical mixtures. *Statistica Sinica*, *13*, 1097–1118.

Jacquier, E., Polson, N., & Rossi, P. (1994). Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business of Econometrics Statistics*, *12*, 371–417.

Kim, S. N., Shephard, N., & Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, *65*, 361–393.

Lopes, H. F., Aguilar, O., & West, M. (2000). Time-varying covariance structures in currency markets. In: *Proceedings of the XXII Brazilian Meeting of Econometrics*, Campinas, Brazil.

Lopes, H. F., & Migon, H. (2002). Comovements and contagion in emergent markets: Stock indexes volatilities. In: C. Gatsonis, A. Carriquiry, A. Gelman, D. Higdon, R.E. Kass, D. Pauler, & I. Verdinelli (Eds), *Case studies in Bayesian statistics*, (Vol. 6, pp. 285–300). New York: Springer-Verlag.

Lopes, H. F., & Salazar, E. (2005). Bayesian inference for smooth transition autoregressive time series models. *Journal of Time Series Analysis* (to appear).

Lubrano, M. (2000). Bayesian analysis of nonlinear time series models with a threshold. In: W. A. Barnett, D. F. Hendry, S. Hylleberg, T. Teräsvirta, D. Tjostheim & A. Würtz (Eds), *Proceedings of the Eleventh International Symposium in Economic Theory*. Helsinki: Cambridge University Press.

Luukkonen, R., Saikkonen, P., & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, *75*, 491–499.

Medeiros, M. C., & Veiga, A. (2005). A flexible coefficient smooth transition time series model. *IEEE Transactions on Neural Networks*, *16*, 97–113.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, *21*, 1087–1091.

Ozaki, T. (1982). The statistical analysis of perturbed limit cycle process using nonlinear time series models. *Journal of Time Series Analysis*, *3*, 29–41.

Robert, C., & Casella, G (1999). Monte Carlo statistical methods. New York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Smith, A., & Gelfand, A. (1992). Bayesian statistics without tears. *American Statistician*, *46*, 84–88.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, *64*, 583–639.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, *89*, 208–218.

Tong, H. (1990). *Non-linear time series: A dynamical systems approach*. Oxford: Oxford University Press.

Xia, Y., & Li, W. K. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, *94*, 1275–1285.

Zhu, L., & Carlin, B. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, *19*, 2265–2278.

APPENDIX:. LSTAR FULL CONDITIONAL DISTRIBUTIONS

In this appendix we provide in detail the steps of our MCMC algorithm for the LSTAR models of Section 2. In what follows, [ξ] denotes the full conditional distribution of ξ conditional on the other parameters and the data. Also, $\sum$ short for $\sum_{t=1}^{n}$. The full conditional distributions are:

[θ] Combining Eq. (4), with $y_t \sim N(z_t'\theta, \sigma^2)$, $(\theta_2|\sigma^2, \gamma) \sim N(0, \sigma^2 e^{\gamma} I_{k+1})$, it is easy to see that $[\theta] \sim N(\tilde{m}_\theta^*, C_\theta^*)$ where $C_\theta^* = (\sum z_t z_t' \sigma^{-2} + \Sigma^{-1})$, $\tilde{m}_\theta^* = \Sigma_\theta(\sum z_t y_t \sigma^{-2})$ and, $\Sigma^{-1} = \text{diag}(0, \sigma^{-2} e^{-\gamma} I_{k+1})$.
[σ^2] From Eq. (4), $\varepsilon_t = y_t - z_t'\theta \sim N(0, \sigma^2)$, which combined with the noninformative prior $\pi(\sigma^2) \propto \sigma^{-2}$, leads to $[\sigma^2] \sim IG((T + k + 1)/2, (e^{\gamma}\theta_2'\theta_2 + \sum \varepsilon_t^2)/2)$.
[γ, c] We sample γ^* and c^*, respectively, from $G[(\gamma^{(i)})^2/\Delta_\gamma, \gamma^{(i)}/\Delta\gamma]$ and $TN(c^{(i)}, \Delta_c)$, a normal truncated at the interval $[c_a, c_b]$, and $\gamma^{(i)}$ and $c^{(i)}$ current values of γ and c.

The pair (γ^*, c^*) is accepted with probability $\alpha = \min\{1, A\}$, where

$$A = \frac{\prod_{t=1}^{T} f_N(\varepsilon_t^*|0, \sigma^2)\, f_N(\theta_2|0, \sigma^2 e^{\gamma^*} I_{k+1})\, \pi(\gamma^*)\pi(c^*)}{\prod_{t=1}^{T} f_N(\varepsilon_t^{(i)}|0, \sigma^2)\, f_N(\theta_2|0, \sigma^2 e^{\gamma^{(i)}} I_{k+1})\, \pi(\gamma^{(i)})\pi(c^{(i)})} \times \frac{\left[\Phi\left(\frac{c_b - c^{(i)}}{\sqrt{\Delta c}}\right) - \Phi\left(\frac{c_a - c^{(i)}}{\sqrt{\Delta c}}\right)\right] f_G(\gamma^{(i)}|(\gamma^*)^2/\Delta_\gamma, \gamma^*/\Delta\gamma)}{\left[\Phi\left(\frac{c_b - c^*}{\sqrt{\Delta c}}\right) - \Phi\left(\frac{c_a - c^*}{\sqrt{\Delta c}}\right)\right] f_G(\gamma^*|(\gamma^{(i)})^2/\Delta_\gamma, \gamma^{(i)}/\Delta\gamma)}$$

for $\varepsilon_t^* = y_t - z_t'(\gamma^*, c^*, y_{t-d})\theta$, $\varepsilon_t^{(i)} = y_t - z_t'(\gamma^{(i)}, c^{(i)}, y_{t-d})\theta$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

ESTIMATING TAYLOR-TYPE RULES: AN UNBALANCED REGRESSION?☆

Pierre L. Siklos and Mark E. Wohar

ABSTRACT

Relying on Clive Granger's many and varied contributions to econometric analysis, this paper considers some of the key econometric considerations involved in estimating Taylor-type rules for US data. We focus on the roles of unit roots, cointegration, structural breaks, and non-linearities to make the case that most existing estimates are based on an unbalanced regression. A variety of estimates reveal that neglected cointegration results in the omission of a necessary error correction term and that Federal Reserve (Fed) reactions during the Greenspan era appear to have been asymmetric. We argue that error correction and non-linearities may be one way to estimate Taylor rules over long samples when the underlying policy regime may have changed significantly.

☆Presented at the 3rd Annual Conference in Econometrics: Econometric Analysis of Financial Time Series, honoring the contributions of Clive Granger and Robert Engle, Louisiana State University, Baton Rouge. Comments on an earlier draft by Øyvind Eitreheim, Bill Gavin, and Charles Goodhart and Conference participants are gratefully acknowledged.

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 239–276
Copyright © 2005 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20029-4

1. INTRODUCTION

How a central bank reacts to changes in economic conditions is of crucial importance in both policy-making and to academics. Attempts to describe a systematic process by which the central bank adjusts a variable that it has control over, such as an interest rate, to changes in economic conditions give rise to the reaction function approach. There are many ways that one can specify monetary policy reaction functions or rules. Several variables have been thought to significantly affect the setting of monetary policy instruments, including monetary aggregates and the exchange rate. In recent years, however, Taylor-type rules have become the most popular way of summarizing how central banks conduct monetary policy. Taylor (1993) showed that US monetary policy, covering a brief period of only 5 years (1987–1992), is well described by movements in the federal funds rate that respond to deviations of inflation and real GDP from their target and potential levels, respectively. Taylor evaluated the response to these two variables and found that the federal funds rate could essentially be described as a rule of the form:

$$i_t = r^* + \pi_t + \alpha(\pi_t - \pi_t^*) + \beta(\tilde{y}_t) \tag{1}$$

where i is the nominal federal funds rate, r^* the equilibrium federal funds rate, π_t the inflation rate over the previous four quarters, π_t^* the target inflation rate, and $\tilde{y}$ the percent deviation of real GDP from its target or potential level. The Federal Reserve (Fed) reacts by changing the federal funds rate in such a manner that i_t is expected to increase when inflation rises above its target or when real GDP rises above its target or potential level.

The above equation can also be rewritten as

$$i_t = \mu + (1 + \alpha)\pi_t + \beta\tilde{y}_t \tag{2}$$

where $\mu = r^* - \alpha\pi^*$. The parameters α and β reflect the preferences of the monetary authority in terms of their attitude toward the short-run trade-off between inflation and output (see Ball, 1997; Muscatelli, Tirelli, & Trecroci, 1999; Clarida, Gali, & Gertler, 1998).[1] In Taylor (1993), r^* and π_t^* are both set equal to 2, and a weight of 0.5 is assigned to both α and β. Stability implies that $\alpha > 0$, which means that the response of the monetary authorities to an inflation shock translates into a higher real federal funds rate. Otherwise, the central bank cannot convince markets that it prefers lower future inflation.[2]

Numerous papers have estimated variants of the above specification. Less thought is given as to whether the right- and left-hand side variables are of the same order of integration. Banerjee, Dolado, Galbraith, and Hendry (1993, pp. 164–166) discuss conditions under which it is preferable to balance regressions where the regressor and the regressand are a mix of $I(1)$ and $I(0)$ series.[3] Sims, Stock, and Watson (1990) show that estimating (2) in levels, even when the series are non-stationary in levels, need not be problematic.[4] However, omission of a possible underlying cointegrating relationship leads to a misspecification with implications for forecast performance (Hendry, 1995, chapter 8; Clements & Hendry, 1993). This is one point Granger (2004) emphasizes in his Nobel Lecture.

While most monetary economists and central bankers would agree on the fundamental features of a monetary policy rule (if they were inclined to use one), there is still disagreement about the details of the specification. In this paper, we are interested in exploring some of the econometric issues that have a bearing on the estimation of Taylor rules, especially ones that stem from the varied seminal contributions of Clive Granger in the field of time series analysis. Most notably, the stationarity and cointegration properties of time series can have a significant impact on the interpretation and forecasts from reaction functions. For example, the real interest rate, r^*, need not be constant as is generally assumed. The difficulty, however, is that the equilibrium real interest rate implied by the rule (r^* in (1)) is unobserved and subject to a great deal of uncertainty (e.g., Laubach & Williams, 2003). Nevertheless, depending upon the chosen sample, it can display trend-like behavior. Alternatively, relying on a standard Fisher equation decomposition, a unit root in either the inflation rate or the real rate will also produce unit root behavior in the nominal rate. Moreover, economic theory suggests that some of the variables in the Taylor rule, which may be non-stationary in levels, might be cointegrated. This suggests a common factor linking these series that gives rise to error correction type models. The nominal interest rate and inflation are candidates for cointegration but other cointegrated relationships are also possible, as we shall see. Perhaps surprisingly, the extant literature has largely ignored this possibility.

In the next two sections, we outline the principal econometric issues in the estimation of Taylor-type equations. Knowing the time series properties of the variables in (1) or (2) is crucial for proper inference. In particular, little thought is given to whether the right- and left-hand side variables are of the same order of integration. The dependent variable in Eq. (2) is treated as a stationary series even though it may be difficult to distinguish it from a unit root process.[5] Next, following a brief review of existing estimates of Taylor

rules for the US, we consider the sensitivity of estimates of Equations such as (1) and (2), and its variants, to the lessons learned from Clive Granger's work. The paper concludes with a summary.

2. SELECTED ECONOMETRIC ASPECTS OF TAYLOR RULES

Our aim is to focus our attention on lessons learned from the contributions of Granger. Nevertheless, a few other issues germane to estimating Taylor rules will be raised though not extensively considered in the subsequent empirical work that follows. Inference from estimates of Equations such as (1) and (2) are sensitive to several econometric considerations. These include:

(i) *The inclusion of a term to accommodate interest rate smoothing behavior of central banks*. The path of interest rates set by central banks tends to be fairly smooth over time, changing slowly in one direction or the other, with reversals occurring relatively infrequently (e.g., Siklos, 2002, Table 4.3). Taylor-type rules have been modified to incorporate interest rate smoothing by including a lagged interest rate term. Sack and Wieland (2000), in a survey, argue that interest rate smoothing may be deliberate or just the result of monetary policy reacting to persistent macroeconomic conditions. If the latter view is correct, one would expect that the coefficient associated with the lagged interest rate to be small and insignificant. However, estimates of Taylor rules find that the coefficient associated with the lagged interest rate is close to unity and statistically significant, suggesting that interest rate smoothing may be deliberate.[6] The conclusion indicating that central banks practice interest rate smoothing has been questioned by Rudebusch (2002) and Söderlind, Söderström, and Vredin (2003). All of them argue that a Taylor rule augmented with a lagged interest rate implies too much predictability of interest rate changes compared with yield curve evidence. Specifically, Rudebusch (2002) argues that a large coefficient on the lagged interest rate would imply that future interest rate changes are highly predictable. But yield curve evidence suggests that interest rate predictability is low. Hence, Rudebusch concludes that the dynamic Taylor rule is a mis-specified representation of monetary policy. Rudebusch is not able to obtain any definitive conclusions based on this test because of an observational equivalence problem affecting the

analysis.[7] Nevertheless, there are solid theoretical and practical reasons to believe that central banks do smooth interest rates (e.g., Goodhart, 1999; Sack, 1998; Sack & Wieland, 2000; Collins & Siklos, 2004; Bernanke, 2004b). Moreover, English et al. (2003) reject on empirical grounds Rudebusch's conjecture. Dueker and Rasche (2004) argue that US studies, which rely on quarterly or monthly data fail to make an adjustment arising out of the discrete nature of changes in the target fed funds rate. Nevertheless, once the appropriate adjustment is made, interest rate smoothing remains a feature of the data.

(ii) *Whether or not the equilibrium real rate is constant*. Taylor-type rules contain both an unobserved equilibrium real interest rate and an unobserved inflation target. Judd and Rudebusch (1998), Kozicki (1999), and Clarida, Gali, and Gertler (2000), among others, calculate the equilibrium real interest rate as the difference between the average federal funds rate and the average inflation rate. With this sample specific value, one is able to back out an estimated value for the inflation target from the empirical estimate of the constant term in Eq. (2) above. Of course, one could also begin with an inflation target and back out a value of the real interest rate from the constant term. Rudebusch (2001) employs a more elaborate approach to estimating the equilibrium real interest rate from empirical estimates of an IS curve. Kozicki (1999) finds that estimates of the equilibrium real interest rate vary over time for the US. Clarida et al. (2000) use the average of the observed real interest rate as a proxy for the equilibrium real rate but allow it to vary between sub-samples. More recently, Rapach and Wohar (2005) have shown that ex-post real interest rates for the G7 countries have undergone structural shifts over time. They find that these shifts correspond to structural breaks in the inflation rate.[8]

(iii) *Estimated weights α and β may be sensitive to the policy regime in place*. Different sample periods may yield different results. Judd and Rudebusch (1998) and Hamalainen (2004) estimate Taylor-type rules for different sample periods and find that the Fed has reacted differently over time to inflation and the output gap. Hence, the possibility of structural breaks, well-known to have a significant impact on the unit root and cointegration properties of the data, also plays a role.

(iv) *Different ways in which the inflation rate can be estimated*. In the original Taylor rule, monetary policy targets inflation as the rate of inflation in the GDP deflator over the previous four quarters. Kozicki (1999) compares the Taylor rule using four different price indexes to

compute the annual inflation rate (GDP price deflator, CPI, core CPI, and expected inflation from private-sector forecasts).

(v) *Different measures of potential output*. The output gap measure used in the Taylor rule has been interpreted as either representing policy-makers objectives relating to output stability as well as a measure of expected inflation (see Favero & Rovelli, 2003). Different methods have been used to construct proxies for potential output. In the original Taylor (1993) paper, potential output was computed by fitting a time trend to real GDP. Judd and Rudebusch (1998) and Clarida et al. (2000) measure potential output using the Congressional Budget Office's (CBO) estimates and by fitting a segmented trend and quadratic trend to real GDP. McCallum and Nelson (1999) and Woodford (2001) have argued against the use of time trends as estimates of potential output. First, because the resulting output gap estimates might be overly sensitive to the chosen sample; second, de-trending ignores the potential impact of permanent shocks on output. The latter consideration proves especially relevant when the time series properties of the output gap series are investigated, as we shall see. Other measures of potential output in the extant literature include calculations using the Hodrick–Prescott (HP) filter, and a band-pass filter. Kozicki (1999) shows that different measures of potential output can lead to large variations in Federal Funds rate target movements. Finally, Walsh (2003a, b) argues that policy makers are best thought of as reacting to the *change* in the output gap. Typically, however, estimated output gap proxies are derived under the implicit, if not explicit, assumption that the series is stationary. Differencing such a series may further exacerbate the unbalanced regression problem noted earlier and may result in over-differencing.[9]

(vi) *The timing of information (e.g., backward or forward looking expectations) as well as the use of current versus real time data*. In Taylor's original article, the Fed reacts contemporaneously to the right-hand side variables. However, this assumes that the central bank knows the current quarter values of real GDP and the price index when setting the federal funds rate for that quarter, which is unlikely.[10] Levin, Wieland, and Williams (1999), McCallum and Nelson (1999), and Rudebusch and Svensson (1999) found few differences between the use of current versus lagged values of variables, perhaps because both inflation and the output gap are very persistent time series so that lagged values serve as good proxies for current values. Clarida et al. (1998, 2000), Orphanides (2001), and Svensson (2003) emphasize the forward-looking

nature of monetary policy and advocate the estimation of forward-looking Taylor-type rules. Orphanides (2001) estimates Taylor-type rules over the period 1987–1992 using ex-post revised data as well as real-time forecasts of output gap. He also used Federal Reserve staff forecasts of inflation to investigate whether forward looking specifications describe policy better than backward-looking specifications. Differences in the interpretation of results based on whether models are forward or backward-looking, forecast-based, or rely on real time data strongly points to a potentially important role of the time series properties of the variables that enter the Taylor rule (also see Tchaidze, 2004).

(vii) *Issues of model uncertainty*. Svensson (2003) doubts the relevance of the Taylor rule on theoretical grounds. Levin et al. (1999) and Taylor (1999) argue that Taylor-type rules are robust to model uncertainty.[11] However, these conclusions concerning robustness under model uncertainty could be challenged on the grounds that the models they use are too similar to each other. A potential solution may be to adopt the "thick modeling" approach of (Granger & Jeon, 2004; also see Castelnuovo & Surico, 2004), which we will not pursue here.[12]

These foregoing issues point to the crucial role played by the time series properties of the nominal interest rate, inflation, and output gap. The interest rate and inflation rate in Eq. (2) have often been found to be non-stationary $I(1)$ processes, while the output gap is usually, by construction, a stationary $I(0)$ process (see Goodfriend, 1991; Crowder & Hoffman, 1996; Culver & Papell, 1997). This would suggest that it might be preferable to estimate a first difference form of the Taylor rule (e.g., as in Siklos, 2004; Gerlach-Kristen, 2003).[13]

Regardless of whether interest rates and inflation are non-stationary or stationary processes, most would agree that these series may be near unit root processes or, rather, that their stationarity property can be regime dependent. There are problems with estimating equations with near unit root processes. Phillips (1986, 1989) has shown that if variables are integrated of order 1, or are near unit root processes, then levels regressions may yield spurious results, and standard inference is not asymptotically valid (Elliott & Stock, 1994). Alternatively, Lanne (1999, 2000) draws attention to the fact that the finding of a unit root in nominal interest rates is common because there is likely a structural break that gives rise to the unit root property. Even if one believes that a unit root in interest rates, or any of the other variables in a Taylor rule, is a convenient simplification, there remains

the problem with the economic interpretation of a unit root in interest rates. There are also good economic reasons to expect some cointegration among the variables in the Taylor rule. For example, the real interest rate in (2) reflects the Fisher equation relating the nominal interest rate to expected inflation, and these two series may be statistically attracted to each other in the long run if real interest rates are believed to be stationary.[14]

If $I(1)$ variables are found not to be cointegrated, a static regression in levels will be spurious (Granger & Newbold, 1974), and in a spurious regression the estimated parameter vector is inconsistent and the t- and F-statistics diverge. Studies of the Taylor rule which have regression equations with R^2 greater than the DW statistic is an indication of spurious regression.[15] If no cointegration is found then there is no long-run relationship between the $I(1)$ variables.

3. ALTERNATIVE SPECIFICATIONS OF THE TAYLOR RULE

Judd and Rudebusch (1998) modify Taylor's rule given in Eq. (2) above that allows for gradual adjustment in the Federal funds rate. The modified Taylor rule is written as

$$i_t^* = r^* + \pi_t + \alpha(\pi_t - \pi^*) + \beta_1 \tilde{y}_t + \beta_2 \tilde{y}_{t-1} \tag{3}$$

where the adjustment process is given by

$$\Delta i_t = \gamma(i_t^* - i_{t-1}) + \rho \Delta i_{t-1} \tag{4}$$

The coefficient γ is the adjustment to the "error" in the interest rate setting and the coefficient ρ can be thought of as a measure of the "momentum" from last period's interest rate change.[16] Combining (3) and (4) gives the Taylor-type rule with interest rate smoothing:

$$\Delta i_t = \gamma\mu - \gamma i_{t-1} + \gamma(1+\alpha)\pi_t + \gamma\beta_1 \tilde{y}_t + \gamma\beta_2 \tilde{y}_{t-1} + \rho \Delta i_{t-1} \tag{5}$$

where $\mu = r^* - \alpha\pi^*$.

Note, however, that (5) is a difference rule and, as Hendry (2004) points out, it was the dissatisfaction with specifications such as these, notably the fact that such a specification would be valid whether the raw data were $I(1)$ or $I(0)$, which eventually found expression in what eventually came to be called Granger's representation theorem.

The above Equation can be rewritten as

$$i_t = \gamma\left[\mu + \gamma(1+\alpha)\pi_t + \beta_1\tilde{y}_t + \beta_2\tilde{y}_{t-1}\right] + (1-\gamma)i_{t-1} + \rho\Delta i_{t-1} \tag{6}$$

Clarida et al. (1998, 2000) and McCallum (2001) employ a partial-adjustment form of the Taylor rule that is further modified to reflect forward-looking behavior of monetary policy makers. This interest rate rule is given as

$$i_t^* = i^* + \varphi(E[\pi_{t+k}|\Omega_t] - \pi^*) + \beta E\left[\tilde{y}_{t+m}|\Omega_t\right] \tag{7}$$

where i_t^* is the nominal interest rate to be set; i^* the desired nominal interest rate when inflation and output are at their target values; π_{t+k} is the inflation rate over the period t to $t+k$; $\tilde{y}_{t+m}$ a measure of the average output gap between period t and $t+m$; Ω_t the information set available at the time the interest rate is set. Interest rate smoothing is incorporated through a partial-adjustment process given as

$$i_t = \rho(L)i_{t-1} + (1-\rho)i_t^* \tag{8}$$

where

$$\rho(L) = \rho_1 + \rho_2 L + \cdots + \rho_n L^{n-1}; \quad \rho \equiv \rho(1)$$

where i_t is the current interest rate set by the monetary authority. Combining (7) and (8) yields the instrument rule to be estimated as

$$i_t = (1-\rho)\left[r^* - (\varphi-1)\pi^* + \varphi\pi_{t+k} + \beta\tilde{y}_{t+m}\right] + \rho(L)i_{t-1} + \varepsilon_t \tag{9}$$

where $r^* = i^* - \pi^*$ is the long-run equilibrium real rate and

$$\varepsilon_t = -(1-\rho)\left[\varphi(\pi_{t+k} - E[\pi_{t+k}|\Omega_t]) + \beta\left(\tilde{y}_{t+m} - E\left[\tilde{y}_{t+m}|\Omega_t\right]\right)\right] \tag{10}$$

Levin et al. (1999) argue that the behavior depicted in Eq. (9) could be an optimal response for a central bank. Such interest rate smoothing has been employed in the empirical work of Carida et al. (1998, 2000), Gerlach and Schnabel (2000), and Domenech, Ledo, and Taguas (2002).

While differencing the variables in the Taylor rule solves one problem it creates another. Castelnuovo (2003) tests for the presence of interest rate smoothing at quarterly frequencies in forward-looking Taylor rules by taking into account potentially important omitted variables such as the squared output gap and the credit spread. Collins and Siklos (2004) estimate optimal Taylor rules and demonstrate empirically that interest rate smoothing is the trade-off for responding little to the output gap. Uncertainty about the measurement of the output gap is well known (e.g., see Orphanides, 2001) but it could just as well be uncertainty about the appropriate structural model of the economy that leads policy members to react cautiously.

While the original Taylor rule was specified as having the Fed reacting to only two variables, more recently several papers have explored the potential role of real exchange rate movements (e.g., as in Leitemo & Røisland, 1999; Medina & Valdés, 1999) or the role of stock prices (e.g., Bernanke & Gertler, 1999; Cecchetti, Genberg, Lipsky, & Wadhwani, 2000; Fuhrer & Tootell, 2004). Lately, there has been interest concerning the impact of the housing sector on monetary policy actions as many central banks have expressed concern over the economic impact of rapidly rising housing prices (e.g., see Siklos & Bohl, 2005a, b and references therein). While it is too early to suggest that there is something of a consensus in the literature, it is clear that variables such as stock prices and housing prices may result in the need to estimate an augmented version of Taylor's rule. In what follows, we do not consider such extensions.

4. EMPIRICAL EVIDENCE ON TAYLOR-TYPE RULES: A BRIEF OVERVIEW OF THE US EXPERIENCE[17]

In Taylor (1993) no formal econometric analysis is carried out. Nevertheless, the simple rule described in Eq. (1) above was found to visually track the federal funds rate for the period 1987–1992 fairly well. Taylor (1999) estimated a modified version of Eq. (1), shown as Eq. (2) above, for different sample periods using OLS. Taylor concluded that the size of the response coefficients had increased over time between the international gold standard era and the Bretton Woods and Post-Bretton Woods period. Using the same specification, Hetzel (2000) examines the sample periods 1965–1979, 1979–1987, and 1987–1999 for the US. He also finds that the response coefficients increased over time, but Hetzel questions whether these could be given any structural interpretation. Orphanides (2001) estimates the Taylor rule for the US over the period 1987–1992 using OLS and IV and employs both revised- and real-time data. He finds that when real-time data are used, the rule provides a less accurate description of policy than when ex-post revised data are used. He also finds that the forward-looking versions of the Taylor rule describe policy better than contemporaneous specifications, especially when real-time data are used.

Clarida et al. (2000) consider two sub-samples 1960–1979 and 1979–1998 for the US using the same specification and employing the General Method of Moments (GMM) estimator. The error term in Eq. (9) is a linear combination of forecast errors and thus, it is orthogonal to the information set Ω_t. Therefore, a set of instruments can be extracted from the information set

to use in the GMM estimation. They report considerable differences in the response coefficients over the two different sample periods.[18] The results of Clarida et al. (2000) are challenged by the findings of Domenech et al. (2002). They find evidence for activist monetary policy in the US in the 1970s that resembles that of the 1980s and 1990s. On balance, however, as documented by Dennis (2003), Favero and Rovelli (2003), and Ozlale (2003), the bulk of the evidence suggests that the Volcker–Greenspan era differs markedly from monetary policy practiced by their predecessors at the Fed.

Hamalainen (2004) presents a brief survey of the properties relating to Taylor-type monetary policy rules focusing specifically on the empirical studies of Taylor (1999) (using the original Taylor rule), Judd and Rudebusch (1998), incorporating interest rate smoothing into a modified Taylor rule, and Clarida et al. (2000), incorporating forward-looking behavior into a modified Taylor rule. Different measures of inflation and output gaps as well as different sample periods are employed. Not surprisingly, perhaps, the response coefficients are sensitive to the measure of inflation and output gap employed and to the sample estimation period employed. Hamalainen (2004) also notes that small changes in assumptions can yield very different policy recommendations, leading one to question the use of Taylor-type rules.[19]

There have only been a small number of papers that have examined the time series properties of the variables within Taylor-type rules.[20] Christensen and Nielsen (2003) reject the traditional Taylor rule as a representation of US monetary policy in favor of an alternative stable long-run cointegrating relationship between the federal funds rate, the unemployment rate, and the long-term interest rate over the period 1988–2002. They find that deviations from this long-run relation are primarily corrected via changes in the federal funds rate.

Osterholm (2003) investigates the properties of the Taylor (1993) rule applied to US, Australian, and Swedish data. The included variables are found to be integrated or near integrated series. Hence, a test for cointegration is carried out. The tests find no support for cointegration though the economic rationale for testing for cointegration among the series in the Taylor rule is not justified. Siklos (2004) reports similar results for a panel of countries that includes the US (also see Ruth, 2004), but whether the Taylor rule represents an unbalanced or spurious regression remains debatable, and one we explore econometrically in greater detail below.

More recently, empirical estimates of Taylor's rule have begun to consider non-linear effects. Siklos (2002, 2004) allows for asymmetric changes in inflation and finds in favor of this view in a panel of OECD countries, as well as for GARCH effects in interest rates. Dolado, Pedrero,

and Ruge-Murcia (2004) also permit ARCH-type effects in inflation (i.e., via the addition of the variance in inflation in the Taylor rule) but find that nonlinear behavior describes US monetary policy well only after 1983.

In what follows, we rely exclusively on OLS estimation since these are adequate for bringing out the types of econometric problems raised in Granger's work and relevant to our understanding and assessment of Taylor rules.

5. TAYLOR RULE ESTIMATES FOR THE US

5.1. Data

The data are quarterly converted via averaging of monthly data.[21] The full sample consists of data covering the period 1959.1–2003.4. The mnemonics shown in parentheses are the labels used to identify each series by FREDII at the Federal Reserve Bank of St. Louis.[22] The short-term interest rate is the effective federal funds rate (FEDFUNDS). To proxy inflation, evaluated as the first log difference of the price level, we examine a number of price indices. They are: the CPI for all urban consumers and all items (CPIAUCSL), the core version of this CPI (CPILFESL; excludes food and energy prices), and the price index for Personal Consumption Expenditures (PCECTPI), as well as the core version of this index (JCXE). All these data were seasonally adjusted at the source. In the case of output, real Gross Domestic Product (GDPC96), and real potential Gross Domestic Product (these are the Congressional Budget Office's estimates; GDPPOT2) were utilized. Again, the real GDP data were seasonally adjusted at the source. We also consider other proxies for output gap including, for example, the difference between the log of output less its HP filtered value.[23]

Some studies use a proxy for the unemployment gap. This is defined as the unemployment rate less an estimate of the NAIRU. This can be estimated as in Staiger, Stock, and Watson (1997).[24] We also relied on other data from time to time to examine the sensitivity of our results to alternative samples or regimes. For example, Romer and Romer's Fed contraction dates (Romer & Romer, 1989; 1947.01–1991.12), Boschen and Mills' indicator of the stance of monetary policy (Boschen & Mills, 1995; 1953.01–1996.01), Bernanke and Mihov's (1998; 1966.01–1996.12) measure of the stance of monetary policy, and the NBER's reference cycle dates (www.nber.org).[25]

Forecast-based versions of Taylor rules have also proved popular both because they capture the forward-looking nature of policy making and permit estimation of the reaction function without resort to generated

regressors. In the present paper we rely on eight different sources of forecasts for inflation and real GDP growth. They are from *The Economist*, Consensus Economics, the OECD, the IMF (World Economic Outlook), the Survey of Professional Forecasters, the Livingston survey regularly updated by the Federal Reserve Bank of Philadeplhia, the Greenbook forecasts, and the University of Michigan survey.

5.2. Alternative Estimates and their Implications

Prior to presenting estimates, it is useful to examine the time series properties of the core series described above. Figure 1A plots the fed funds rate while Fig. 1B plots various measures of inflation. Both these figures show a general upward trend until the late 1970s or early 1980s followed by a general downward trend thereafter. Both series appear to be stationary, but only after 1990 approximately. Nevertheless, it is not difficult to conclude, as have several authors, that both these variables are non-stationary over a longer sample that includes a period of rising rates and inflation versus the era of falling rates and disinflation. This is also true even if the sample is roughly split between the pre-Greenspan era (1987) and the period since his appointment as the Fed Chairman.[26] Summary statistics and some unit root tests that broadly support the foregoing discussion are shown in Table 1A. Both the widely-used Augmented Dickey–Fuller test statistic and the statistically more powerful modified unit root tests proposed by Elliott, Rothenberg, and Stock (Ng & Perron, 2001) generally lead to the same conclusion.

Taylor rule estimates based on available forecasts generally do not consider the time series properties of the variables in question. Figure 2A plots actual CPI inflation against forecasted inflation from eight different sources. Since data are not available for the full sample from all sources, only the sample since Greenspan has been Fed Chairman is shown. The top portion of the figure collects forecasts that are survey based in addition to those from the Fed's Greenbook. The bottom portion of the figure contrasts forecasts for inflation from international organizations or institutions that collect and average forecasts from several contributors or sources. Survey-based forecasts appear to be far more volatile than the other types of forecasts. Otherwise, neither type of forecast appears outwardly to dominate others over time. Turning to unit root tests shown in Table 1B, rejections of the null of a unit root in inflation forecasts are more frequent than for realized inflation. This is true for both the full sample as well as the

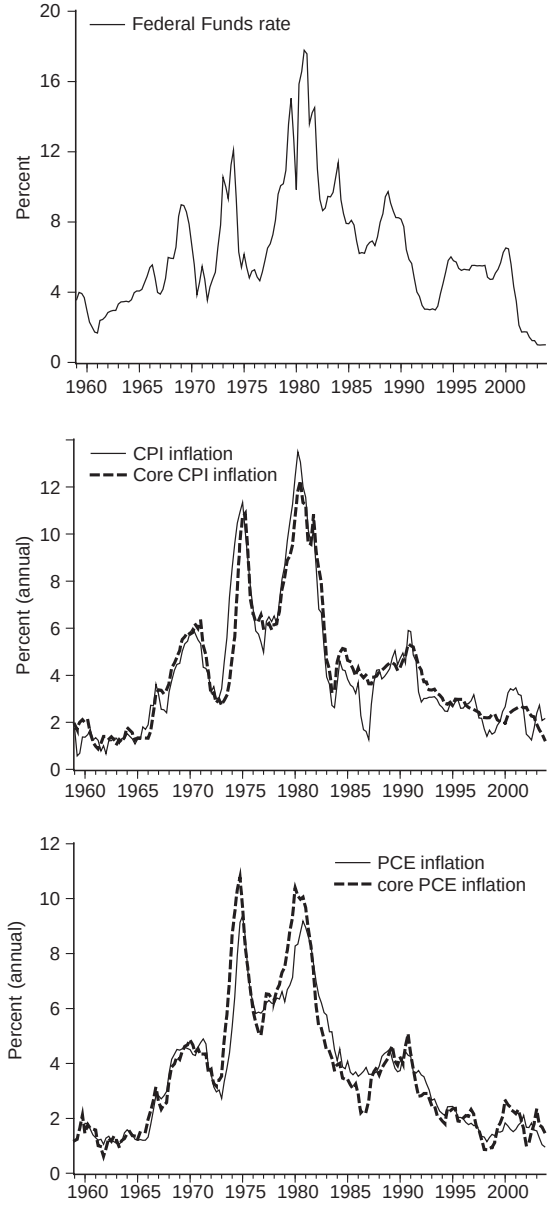


Fig. 1. (A) Federal Fund Rate, 1959:1–2003:4. (B) Alternative Measures of US Inflation. *Source:* See Text.

Table 1A. Summary Statistics and Unit Root Tests.

Series	Sample	Mean	Standard Deviation	ADF *ERS(modified)*
i	Full	6.26	3.32	−1.89 (0.34) ***3.98(7)***
π_{cpi}		4.18	2.82	−1.70 (0.43) *9.70(12)*
$\pi_{cpi-core}$		4.22	2.56	−1.35 (0.60) *11.11(13)*
π_{pce}		3.65	2.15	−1.40 (0.58) *11.29(12)*
$\pi_{pce-core}$		3.73	2.43	−1.55 (0.51) *8.58(13)*
$\tilde{y}_{cbo}$		−1.29	2.59	−3.46 (0.01) *1.59(1)*
$\tilde{y}_{HP}$		−0.03	1.55	−5.85 (0.00) ***1.18(0)***
$\tilde{y}$		−0.01	2.16	−4.16 (0.001) *11.96(12)*
$\Delta\tilde{y}$		0.003	0.86	−7.07 (0.00) *13.94(9)*
$(y-y^{T})$		0.107	3.09	−2.96 (0.04) *4.83(2)*
$\tilde{y}^{T3}$		0.000	2.38	−3.94(0.00) *1.61(1)*
i	Pre-Greenspan	7.05	3.61	−3.52 (0.04) *8.55(2)*
π_{cpi}		4.85	3.29	−1.07 (0.93) *20.44(12)*
$\pi_{cpi-core}$		4.88	2.96	−0.45 (0.98) *26.69(12)*
π_{pce}		4.35	2.33	−1.73 (0.73) *17.14(8)*
$\pi_{pce-core}$		4.46	2.71	−0.82 (0.96) *18.36(12)*
$\tilde{y}_{cbo}$		−1.67	2.94	−2.73 (0.07) *7.80(1)*
$\tilde{y}_{HP}$		−0.08	1.81	−4.28 (0.00) *6.13(0)*
$\tilde{y}$		−0.004	2.51	−2.80 (0.06) *10.44(12)*
$\Delta\tilde{y}$		0.004	1.01	−8.49 (0.00) *21.78(10)*
$(y-y^{T})$		0.258	3.78	−2.34 (0.16) *16.86(1)*
$\tilde{y}^{T3}$		−0.258	2.61	−3.20(0.02) *6.29(1)*
i	Greenspan	4.95	2.26	−2.73 (0.23) *5.91(1)*
π_{cpi}		3.07	1.09	−2.62 (0.27) *9.12(4)*
$\pi_{cpi-core}$		3.13	1.01	−3.56 (0.04) *14.50(4)*
π_{pce}		2.49	1.09	−3.35 (0.07) *13.80(5)*
$\pi_{pce-core}$		2.51	1.08	−2.09 (0.54) *14.06(8)*
$\tilde{y}_{cbo}$		−0.65	1.71	−2.52 (0.12) *16.11(1)*
$\tilde{y}_{HP}$		0.05	0.99	−3.46 (0.01) *11.66(0)*
$\tilde{y}$		−0.03	1.42	−3.29 (0.02) *6.40(4)*
$\Delta\tilde{y}$		−0.001	0.54	−3.40 (0.01) *5.79(1)*
$(y-y^{T})$		−0.15	1.29	−2.30 (0.17) *10.50(0)*
$\tilde{y}^{T3}$		0.45	1.85	−2.43 (0.14) *15.85(1)*

Note: The full sample is 1959.1–2003.4, Pre-Greenspan is 1959.1–1987.2, Greenspan is 1987.3–2003.4. ADF is the Augmented Dickey–Fuller statistic with lag augmentation selected according to the Akaike Information criterion. A constant only is used in the full sample; for the interest rate and inflation series, a constant and a trend is used in the other two samples. *p*- values are given in parentheses. Test statistics in italics are based on the modified ERS tests due to Ng and Perron (2001) using the modified AIC criterion. When the two unit root tests contradict each other this is indicated in **bold**. Lag length in the augmentation portion of the test equation shown in parenthesis. i = interest rate; π = inflation; $\tilde{y}$ = output gap; HP = HP filter; cbo = Congressional Budget Office; pce = Personal Consumption Expenditures; y^{T} is trend output. The figures in brackets are based on a one-sided HP filter.

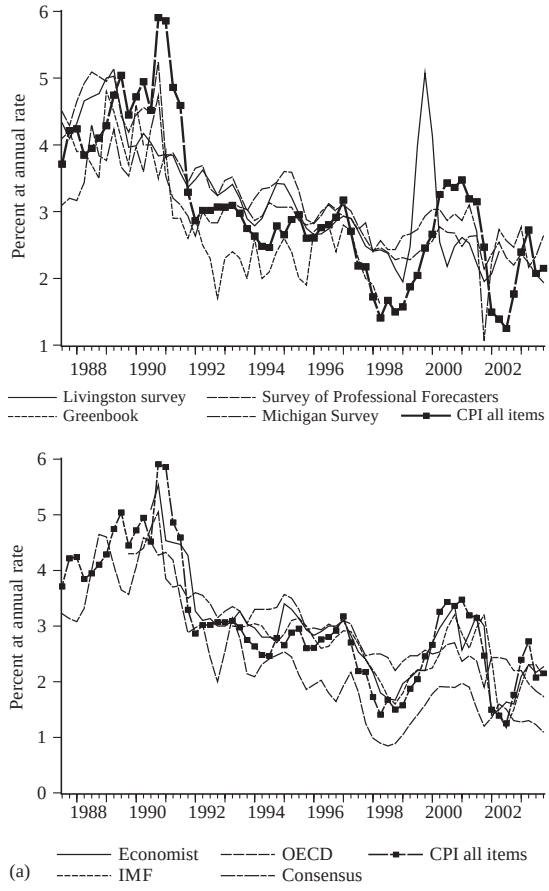


Fig. 2(A). Alternative Forecasts of Inflation: 1987–2003. *Source:* See Text.

Greenspan era. When output growth forecasts are considered there is more unit root like behavior in the full sample than in the various output gap proxies considered in Table 1A, while the evidence of unit root behavior is just as mixed in the forecast as in the actual data. It is surprising that there have been relatively few Taylor rule estimates that rely on publicly available forecasts. After all, as then Governor Bernanke noted, such forecasts are important to policymakers since "…the public does not know but instead must estimate the central bank's reaction function and other economic relationships using observed data, we have no guarantee that the economy will

Table 1B. Additional Unit Root Tests: Forecast of Inflation and Output Growth.

Source of Forecast	Full Sample	Pre-Greenspan	Greenspan
Inflation			
Livingston	−1.66 (0.45)/*24.78(11)*	−1.40(0.58)/*23.10(8)*	−3.23(0.09)T/*31.18(8)*
Greenbook	−3.77(0.02)T/*18.81(11)*	−2.83(0.19)T/*6.90(0)*	−3.66(0.04)T/*25.66(3)*
Survey of Prof. For.	−4.48(0.00)T/*13.74(0)*		−3.61(0.04)T/***5.85(1)***
Univ. Michigan	−4.70(0.00)T/*9.55(0)*	−2.58(0.11)/*7.30(5)*	−3.68(0.03)T/*7.11(2)*
Economist			−3.74(0.03)T/*11.89(0)*
IMF			−1.53(0.51)/*7.35(6)*
OECD	−1.37(0.59)/*8.29(11)*		−2.59(0.29)T/*6.80(10)*
Consensus			−3.34(0.07)T/*5.92(0)*
Output Growth			
Greenbook	−4.84(0.00)/*6.40(11)*		−2.94(0.05)/*3.63(2)*
Economist			−2.31(0.17)/*13.13(3)*
OECD	−3.57(0.01)/*8.24(12)*		−2.12(0.24)/*1.33(6)*
Consensus			−4.46(0.00)/*13.13(8)*
Livingston	−1.90(0.33)/*2.11(13)*		−2.91(0.17)/*13.05(8)*
Survey of Prof. For.	−1.26(0.65)/*31.91(0)*		−3.45(0.05)/*7.07(0)*

Note: Test based on the ADF test with lag augmentation chosen as in Table 1(A); the test statistics in parenthesis is the modified ERS test as described in Table 1(A). "T" indicates that a deterministic trend was added to the test equation. Blanks indicate insufficient or no data to estimate the test equation. *p*-values are in parenthesis. OECD and Greenbook inflation forecasts are for the GDP deflator and OECD output growth forecasts are for the OECD's estimate of the output gap.

converge – even in infinite time – to the optimal rational expectations equilibrium (Bernanke, 2004a, p. 3). Since the form of the estimated Taylor rule will depend in part on the unit root property of inflation, differences in the time series properties of realized versus forecasted inflation may be important.

Clive Granger's work has long focused on the ability of econometric models to forecast time series. An influential contribution was Bates and Granger (1969) in which some linear combination of forecasts often outperforms even the best forecasts. This result, seemingly counter-intuitive, comes about simply because averaging forecasts is a useful device in canceling out biases in individual forecasts. Figure 2B plots a variety of inflation forecasts. Data limitations govern the chosen samples. As we shall see, averaging inflation forecasts results in estimates of forecast-based reaction functions that compare well with Taylor rules based on realized data. This

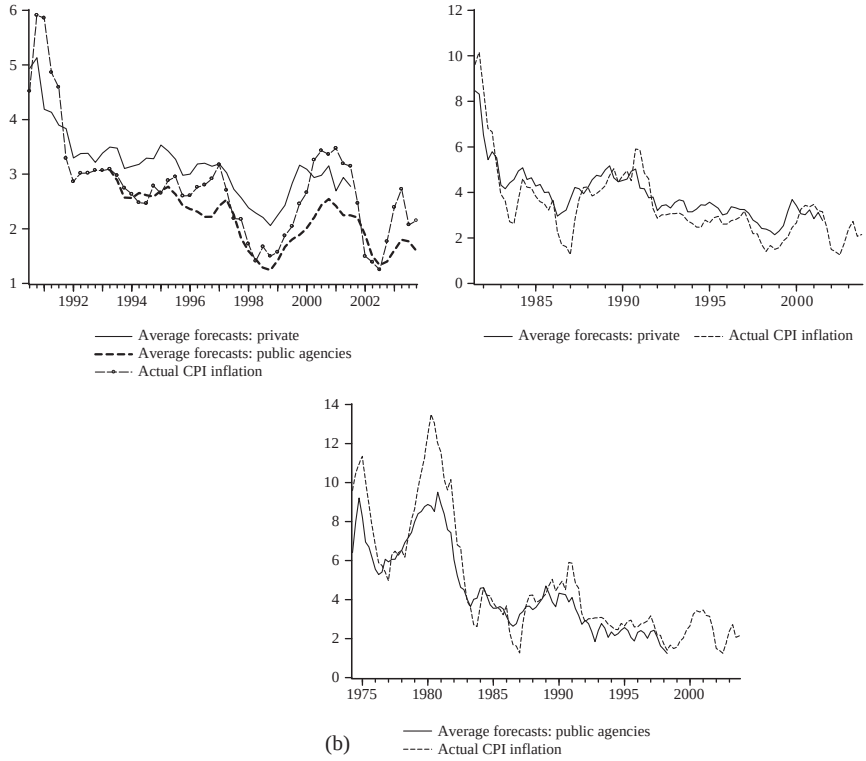


Fig. 2(B). Average Inflation Forecasts of Inflation: 1975–2002. *Note:* In the Top Left Figure, inflation Forecasts from the Greenbook, IMF, and OECD (Public), Survey of Professional Forecasters (SPF), Livingston, Economist (ECON), University of Michigan (UMICH), Consensus (CONS); the Top Right Plot is an Average of SPF, Livingston, and UMICH forecasts; the Bottom Plot is an Average of IMF and OECD Forecasts. Greenbook and OECD Forecasts are for the Implicit Price Deflator.

result is not generally obtained when policy rule are estimated with individual inflation forecasts.

Figures 3A and B display various plots of output growth and the output gap. The time series properties of these series differ noticeably from the other core variables in the Taylor rule. In the top portion of Fig. 3A we compare the output gap first by taking the difference between the growth of actual real GDP and the growth in potential real GDP and next by simply taking the difference between the log levels of actual and potential real GDP. Potential real GDP estimates are from the CBO. The latter is often the

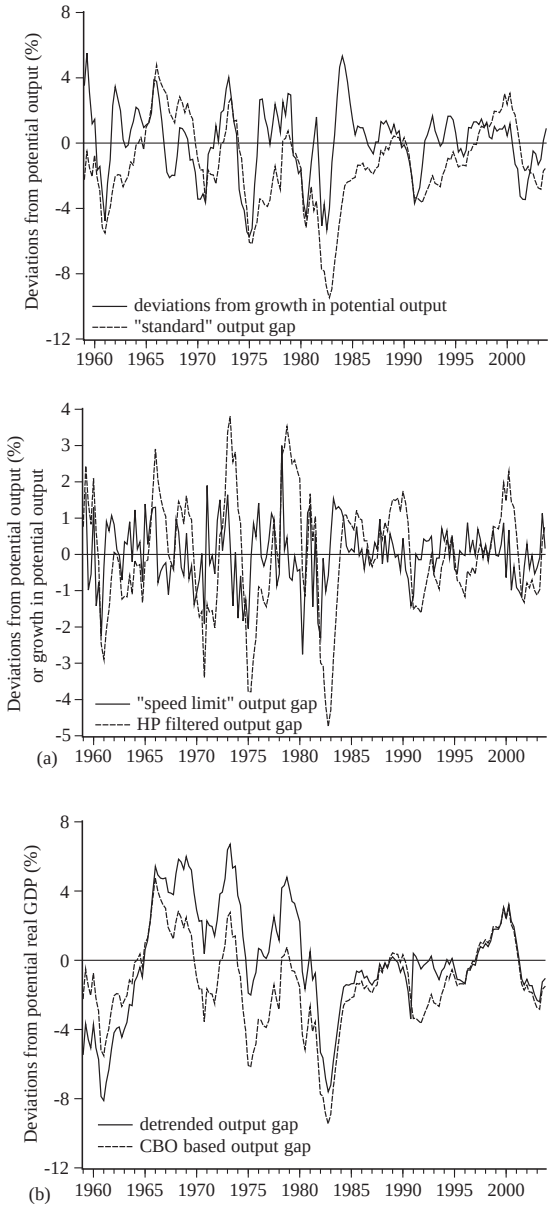


Fig. 3. (A) Alternative Proxies for the Output Gap. *Note:* The "Standard" Output Gap Measure Relies on CBO Estimates of Potential Real GDP. (B) Comparing Detrended and CBO Measures of the Output Gap. *Source:* See Text.

"standard" way of representing the output gap in the empirical literature that relies on US data. The former is a frequently used proxy for the output gap. While perhaps more volatile, the resulting series appears more nearly stationary. The bottom portion of Fig. 3A contrasts a measure of the output gap evaluated as the difference between the log level of real GDP and its HP filtered value against the first difference in the "standard" output gap measure shown in levels in the top portion of the same figure. The latter is sometimes referred to as the "speed limit" measure of the output gap.

Both series are, broadly speaking, stationary. The HP filter suffers from the well-known end point problem. Yet, even if one estimates an output gap measure based on a one-sided HP filter where the end point is permitted to change over time, this does not seem to affect the unit root property of the resulting time series, with the possible exception of data covering the Greenspan era only (see Table 1A). The speed limit measure is clearly more volatile and more stationary in appearance. Finally, Fig. 3B compares the output gap derived from the CBO's measure (i.e., the "standard" proxy shown in Fig. 3A) against a de-trended measure of the log level of real GDP. Versions of the latter have been used in some of the studies cited above. In the version shown here, separate trends for the 1959–1990 and 1991–2003 periods were fitted to the log level of real GDP. The time period captures the phenomenon noted by several analysts who have argued the US experienced a permanent growth in potential output beginning around the 1990s as the impact of efficiencies in computing technology became economically meaningful. Chairman Greenspan has, on a number of occasions, commented on the apparent technological driven boom of the 1990s (based, in part, on evidence as in Oliner & Sichel, 2000).[27] Although the two series produce broadly similar patterns, the measure of the output gap derived from detrending comes closest to being $I(1)$ of all the proxies considered. The same result holds for the Greenspan sample when a cubic trend serves to proxy potential real GDP ($\tilde{y}^{T3}$). If one were instead to rely on the unemployment rate (results not shown) there is a slightly better chance of concluding that the unemployment rate gap is $I(1)$, at least for the separate sub-samples 1957–1987 and 1988–2003, but not when an HP filter is applied.

Notwithstanding the low power of unit root tests, and the presence of structural breaks in the data, perhaps due to regime changes in monetary policy, there are strong indications that interest rates and inflation are $I(1)$ while the output gap is $I(0)$. Nevertheless, since the output gap (or, for that matter, the unemployment rate gap) is a theoretical construct there is no reason that it cannot be $I(1)$ for some sub-sample. In any event, there is

prima facie evidence that the Taylor rule is, as usually estimated, an unbalanced regression.

The possibility that Eq. (2) contains two or more $I(1)$ series raises the possibility that they are cointegrated. Indeed, if one examines Eq. (3) and adds an interest rate smoothing parameter, then it is a cointegrating relationship when $|\rho| < 1$. Alternatively, there may be other cointegrating relationships implicit in the standard formulation of Taylor's rule.[28] One must also ask what economic theory might lead these series to be attracted to each other in a statistical sense. Clearly, the nominal interest rate and inflation are linked to each other via the Fisher effect. It is less clear whether the output gap will be related to the interest rate in the long-run not only because it is typically $I(0)$ by construction but also because we know that a central bank's willingness to control inflation is regime specific.

Table 2 presents a series of cointegration tests and estimates of the cointegration vectors.[29] We first consider cointegration between the output gap, inflation (in the CPI alone), the fed funds rate and a long-term interest rate. Addition of the long-term interest rate series, proxied here by the yield on 10-year Treasury bonds (GS10), reflects the possibility that the term spread, often used as a predictor of future inflation, short-term interest rates, or output, affects the determination of the fed funds rate. The results in Table 1A and 1B indicate that only the output gap series based on detrending is $I(1)$ and thus, we rely on this series alone in the tests that follow. Not surprisingly, we are able to conclude that the output gap does not belong in the cointegrating relationship. Relying on the full sample, the null that the output gap is zero in the cointegrating vector cannot be rejected. Moreover, if we estimate a vector error correction model the null that the fed funds rate is weakly exogenous also cannot be rejected.[30] Both sets of results appear consistent with the notion that the output gap is not an attractor in the specified VAR and that the form of the Taylor rule with the fed funds rate reacting to the inflation and output gap is adequate. One cointegrating vector is found in both models considered. Since the finding of cointegration is clearly sensitive to the output gap definition, we instead investigate whether there is cointegration between inflation, and the short and long rates (Model 2). While cointegration ordinarily requires a long span of data, it is conceivable that the sought after property is a feature only of some well-chosen sub-sample. The results in Table 2, at least based on the model that excludes the output gap, suggest that there is cointegration during the Greenspan and full samples. Figure 4 plots the error corrections from the estimated cointegrating vector based on the full and Greenspan samples. Only in the Greenspan sample do the error corrections appear to

Table 2. Cointegration Tests.

Model 1 $[\tilde{y}^T, \pi_t, i_t, i_t^L]$	Cointegration Test Statistics		
	1959.1–2003.4		1987.3–2003.4
No. of cointegrating vectors	Full {9}	Pre-Greenspan {9}	Greenspan {4}
0	32.29 (0.02)	29.59 (0.04)	44.70 (0.00)
1	19.90 (0.10)	22.30 (0.53)	6.98 (0.98)
2	9.55 (0.38)	15.89 (0.20)	5.55 (0.84)
3	7.72 (0.09)	9.16 (0.18)	9.16 (0.31)
Model 2 $[\pi_t, i_t, i_t^L]$	{9}	{12}	{3}
0	29 (0.00)	19.57 (0.12)	26.22 (0.01)
1	15.19 (0.06)	17.21 (0.03)	11.27 (0.23)
2	5.05 (0.28)	3.92 (0.92)	2.00 (0.78)
Cointegrating equation	$\pi_t - 3.53i_t^L + 2.53i +$ 5.15 (0.61)* (0.53)* (1.51)*		$\pi_t - 0.87i_t^L + 0.26i +$ 1.40 (0.09)* (0.07)* (0.41)*

Note: i^L is the long-term interest rate. Johansen's λ_{max} test statistic reported (p-values due to MacKinnon, Haug, & Michelis, 1999, in parenthesis). Lag length for VARs, shown in brackets, chosen on the basis of the likelihood ratio test except for Model 1, Greenspan sample, where the Final Prediction Error criterion was used. The cointegrating equation is expressed in the form shown as in Eq. (13) with standard errors in parenthesis.
*Indicates statistically significant at the 5% level.

be roughly stationary. In the case of the full sample estimates, it is apparent that the series tends to behave asymmetrically. In contrast, the error corrections appear more symmetric when the Greenspan sample alone is considered. Based on the usual testing, it is difficult to know whether it is preferable to estimate a relationship covering the full sample or some portion of it. As noted previously, there are good reasons to think that the full sample may indeed consist of periods where cointegration is switched on or off depending upon the policy regime in place (Siklos & Granger, 1997). Part of the answer depends on the objectives of the study. Clearly, there is some evidence summarized above suggesting that the Taylor rule is a useful way of studying monetary policy over long periods of time, especially for the US experience (also see Orphanides, 2003a, b). If the purpose is to understand the behavior of a central bank during a well defined policy regime then a sub-sample is clearly preferable. Alternatively, if the point that needs to be

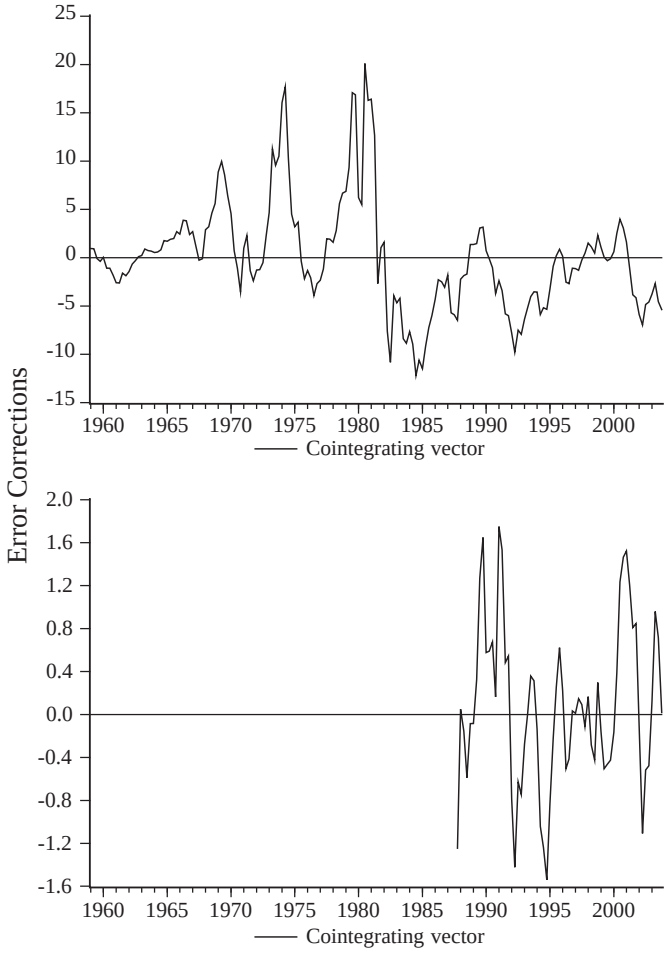


Fig. 4. Error Corrections. *Note:* Obtained from Table 2 Estimates for Model 2 for the Full Sample (top) and Greenspan Samples (bottom).

made is that monetary policy behaves asymmetrically over time then a longer span of data is clearly preferable. The difficulty is in properly estimating the timing of a regime change. As we have seen above there is no widespread consensus on the precise dating of shifts in regimes. Nevertheless, to maximize comparability with other studies, we retain sub-sample distinctions that reflect the impact of the tenures of Volcker and Greenspan.

More sophisticated analyses of the dating of regime changes, as in Rapach and Wohar (2005) for the real interest rate, or Burdekin and Siklos (1999) for inflation, might well choose a slightly different set of dates. However, the conclusion that distinct policy regimes are a feature of US monetary policy over the sample considered would remain unchanged.

We now turn to estimates of the Taylor rule that capture some of the features of the data discussed above. Table 3 provides the main results. Three versions of the Taylor rule were estimated. The first is (9) setting k and m both to zero yielding:

$$i_t = (1-\rho)\left[r^* - (\varphi - 1)\pi^* + \varphi\pi_t + \beta\tilde{y}_t\right] + \rho(L)i_{t-1} + \varepsilon_t \tag{11}$$

The second version recognizes the cointegration property described above. In this case we estimate

$$\Delta i_t = \sum_{i=0}^{\omega}(\varphi_i'\Delta\pi_{t-i} + \beta_i'\Delta\tilde{y}_{t-i} + \rho\Delta i_{t-i}^L) + \delta(\pi - \psi_1 i - \psi_2 i^L)_{t-1} + \rho\Delta i_{t-1} + \zeta_t \tag{12}$$

where all the variables have been defined and $(\pi - \psi_1 i - \psi_2 i^L)_{t-1}$ is the error correction term.[31] A third version decomposes the error correction term into positive and negative changes to capture a possible asymmetry in the reaction function based on the momentum model. The resulting specification is one of a variety of non-linear forces for the error correction mechanism. That specification is written

$$\begin{aligned}\Delta i_t = &\sum(\varphi_i'\Delta\pi_{t-i} + \beta_i'\Delta\tilde{y}_{t-i} + \rho\Delta i_{t-i}^L) + \delta\Delta^+(\pi - \psi_1 i - \psi_2 i^L)_{t-1} \\ &+ \delta\Delta^-(\pi - \psi_1 i - \psi_2 i^L)_{t-1} + \rho\Delta i_{t-1}\zeta_t\end{aligned} \tag{13}$$

The first four columns of Table 3 report the steady-state coefficient estimates for the equilibrium real interest rate, the inflation and output gap parameters as well as the interest rate smoothing coefficient when the estimated model ignores cointegration. It is apparent that real interest rates are relatively lower in the Greenspan period than at any other time since the 1960s. Nevertheless, it is also true that estimates vary widely not only across time but across techniques used to proxy the output gap.[32]

Estimates of the interest rate response to the inflation gap are generally not significantly different from one, the norm known as Taylor's principle. It is clear, however, that the Greenspan era stands in sharp contrast with other samples considered, as the inflation gap coefficient is considerably larger than at any other time reflecting more aggressive Fed reactions to inflation shocks. While the sample here is considerably longer with the

Table 3. Taylor Rule Coefficient Estimates.

	Standard Taylor Rule				Difference Rules		
Sample/ Output gap	$r^*/(1-\rho)$: Real Rate	$\varphi/(1-\rho)$: Inflation	$\beta/(1-\rho)$: Output Gap	ρ: Interest Smoothing	δ: Error Correction	$\Delta\delta^+$: Pos. Momentum	$\Delta\delta^-$: Neg. Momentum
Pre-Volcker							
$\tilde{y}^{T3}$	5.90 (0.02)	0.91 (0.18)	1.40 (0.15)	0.91 (9.99)	−0.09 (−2.41)		
$\Delta\tilde{y}$	0.81 (0.88)	1.04 (0.88)	−6.09 (0.49)	1.04 (17.89)	n.a.		
$\tilde{y}$	5.41 (0.26)	0.97 (0.85)	4.74 (0.56)	0.97 (15.93)	n.a.		
$y-y^{\mathrm{T}}$	8.21 (0.17)	0.94 (0.33)	1.33 (0.38)	0.94 (12.22)	−0.09 (−2.42)		
$\tilde{y}_{\mathrm{HP}}$	5.03 (0.00)	0.82 (0.01)	1.52 (0.00)	0.82 (8.65)	−0.09 (−2.45)		
Volcker							
$\tilde{y}^{T3}$	6.68 (0.00)	0.79 (0.17)	−0.09 (0.67)	0.50 (3.33)	−0.001 (−0.03)		
$\Delta\tilde{y}$	6.48 (0.00)	0.90 (0.41)	0.95 (0.05)	0.38 (2.44)	n.a.		
$\tilde{y}$	6.47 (0.00)	0.90 (0.50)	0.28 (0.19)	0.45 (2.86)	n.a.		
$y-y^{\mathrm{T}}$	6.56 (0.00)	0.82 (0.24)	−0.08 (0.73)	0.51 (3.38)	−0.001 (−0.02)		
$\tilde{y}_{\mathrm{HP}}$	6.73 (0.00)	0.81 (0.20)	−0.09 (0.77)	0.52 (3.44)	0.002 (0.06)		
Greenspan							
$\tilde{y}^{T3}$	2.96 (0.18)	3.80 (0.52)	−0.06 (0.97)	1.05 (16.87)	−0.03 (−1.93)	−.05 (−2.61)	−0.01 (−0.60)
$\Delta\tilde{y}$	5.09 (0.27)	4.31 (0.54)	−19.32 (0.56)	1.02 (32.71)	n.a.	n.n.	
$\tilde{y}$	3.02 (0.04)	0.37 (0.60)	4.29 (0.09)	0.95 (27.66)	n.a.	n.n.	
$y-y^{\mathrm{T}}$	3.12 (0.02)	3.25 (0.13)	0.59 (0.44)	1.06 (27.83)	−0.02 (−1.75)	−0.04 (−2.02)	−0.01 (−0.79)
$\tilde{y}_{\mathrm{HP}}$	2.75 (0.27)	4.39 (0.40)	−0.74 (0.78)	1.04 (24.05)	−0.02 (−1.72)	−0.05 (−2.26)	−0.01 (−0.52)
ỹ-Greenbook	5.12 (0.13)	0.52 (0.65)	1.22 (0.06)	0.89(10.00)			
ỹ-Consensus	1.12 (0.78)	0.88 (0.94)	0.88 (0.03)	0.91(10.36)			

Table 3. (*Continued*)

Sample/ Output gap	Standard Taylor Rule				Difference Rules		
	$r^*/(1-\rho)$: Real Rate	$\varphi/(1-\rho)$: Inflation	$\beta/(1-\rho)$: Output Gap	ρ: Interest Smoothing	δ: Error Correction	$\Delta\delta^+$: Pos. Momentum	$\Delta\delta^-$: Neg. Momentum
Full							
$\tilde{y}^{T3}$	4.60 (0.00)	0.92 (0.34)	0.83 (0.42)	0.91 (25.91)	−0.03 (−1.93)		
$\Delta\tilde{y}$	4.10 (0.00)	1.09 (0.80)	3.30 (0.04)	0.92 (27.11)	n.a.		
$\tilde{y}$	3.63 (0.00)	1.14 (0.56)	1.38 (0.00)	0.88 (25.21)	n.a.		
$y-y^{\mathrm{T}}$	5.27 (0.00)	0.37 (0.18)	0.75 (0.14)	0.93 (27.67)	−0.02 (−1.75)		
$\tilde{y}_{\mathrm{HP}}$	4.71 (0.00)	0.66 (0.20)	1.34 (0.01)	0.90 (24.53)	−0.02 (−1.72)		
Forecast-based rules							
Public-$\tilde{y}_{\mathrm{HP}}$	2.50 (0.07)	1.02 (0.43)	0.12 (0.47)	0.87 (17.40)			
Private-$\tilde{y}_{\mathrm{HP}}$	1.36 (0.10)	0.95 (1.16)	0.29 (0.60)	0.92 (15.33)			
Mix I-$\tilde{y}_{\mathrm{HP}}$	2.54 (0.07)	0.98 (0.32)	0.05 (0.03)	0.88 (22.00)			
Mix II-$\tilde{y}_{\mathrm{HP}}$	3.30 (0.07)	0.89 (0.36)	0.30 (0.26)	0.86(17.20)			

Note: See text and Table 1 for variable definitions. Pre-Volcker = 1959.1–1979.2; Volcker = 1979.3–1987.2; Greenspan = 1987.3–2003.4; Full = 1959.1–2003.4. Equations are estimated via OLS. In parenthesis, p-value for Wald test (F-statistic) that $i^*/(1-\rho)=0$, $\varphi/(1-\rho)=1$, and $\beta/(1-\rho)=0$. For ρ, δ, $\Delta\delta^+$, $\Delta\delta^-$ t-statistic in parenthesis. δ is the coefficient on the error correction term in the first difference form of the Taylor rule. $\Delta\delta^+$, $\Delta\delta^-$ are the coefficient estimates for the positive and negative changes in the error correction term based on the first difference form of the Taylor rule. Greenbook and Consensus are forecast rule estimates. Estimates based on average forecast-based rules are for 1974q2–1998q2 (Public); 1981q3–2001q3 (Private); 1960q1–2002q2 (Mix I); 1960q4–1987q2 (Mix II). Mix I and II represent average forecasts of inflation obtained from the Livingston and OECD sources. Other forecast combinations are described in Fig. 2(B).

addition of 6 years of data, the coefficients reported in Table 3 are markedly higher than the ones reported by Kozicki (1999, Table 3) except perhaps when the conventional output gap using CBO potential real GDP estimates are used.

Turning to the output gap coefficient it is quickly apparent, as others have found (e.g., Favero & Rovelli, 2003; Collins & Siklos, 2004), that the output gap is often insignificant and rather imprecisely estimated. Kozicki (1999, Table 3) reaches much the same conclusion. It is only for the full sample that a non-negligible Fed response to the output gap is recorded. Estimates of the interest rate smoothing parameter confirm the largely close to unit root behavior of nominal interest rates with the notable exception of the Volcker era when interest rate persistence is considerably lower. This is to be expected not only because of the sharp disinflationary period covered by this sample but also because the Fed's operating procedure at the time emphasized monetary control over the interest rate as the instrument of monetary policy.

We also considered estimating various versions of inflation forecast-based Taylor rules. Since the bulk of the available data are only available since the late 1980s we show only some estimates for the Greenspan era.[33] Of eight potential versions of Taylor's rule only two, namely a version that relies on the Greenbook forecasts (of GDP deflator inflation) and the Consensus forecasts produce plausible results.[34] Otherwise, estimates of all the coefficients are implausibly large and the interest rate smoothing coefficient is more often than not greater than one.[35] It is not entirely clear why such a result is obtained but it is heartening that, since both these forecasts are among the most widely watched, forecasters see Fed behavior much the same way as conventional regression estimates based on actual data. Nevertheless, it should be noted that the plausibility of forecast-based estimates is highly sensitive to the choice of the output gap measure. The CBO's output gap measure is the only one that yields results such as the ones shown in Table 3. Matters are different when forecasts are averaged. Whether public agencies or private sector forecasts are averaged no longer appears to matter. The resulting estimates are now comparable to ones that rely on actual data. Another one of Clive Granger's insights proves useful.

The last set of columns considers the significance of adding an error correction term to the Taylor rule in first differences. The error correction term was only added in cases where one could reasonably conclude that $\tilde{y} \sim I(1)$ based on earlier reported tests. It is striking that, with the exception of the Volcker era, the error correction term is significant indicating, as argued above, that long-run equilibrium conditions implicit in Taylor's rule,

should not be omitted. The Volcker exception makes the case for cointegration having been turned off during that period, and this interpretation is unlikely to be a controversial one. Finally, the last two columns of Table 3 give estimates of the error correction terms when a momentum type asymmetric model is estimated (Eq. (13)). Given that asymmetry in monetary policy actions is most likely to have occurred during Greenspan's tenure, we only consider estimates for the post-1987 sample. The results suggest that since only $\Delta\delta^{+}$ is statistically significant this is akin to an interpretation whereby a rise in the term spread, an indicator of higher expected inflation, over inflation (i.e., a negative error correction) prompts the Fed to raise interest rates. The effect does not appear to work in reverse. Therefore, this seems consistent with aggressive Fed policy actions to stem higher future inflation that marks Greenspan's term as Fed Chair.[36] Clearly, there are other forms of non-linearity that could have been employed. For example, as shown in Table 4, there is evidence of ARCH-type effects in the residuals of the Taylor rule in every sample although the evidence is perhaps less strong for the Greenspan era. Obviously, a task for future research is to provide not only a more rigorous theory that might lead to a non-linear rule but more extensive empirical evidence.

As a final illustration of the usefulness of the error correction term we examine in-sample forecasts of the Fed funds rate during the Greenspan term. Figure 5A plots forecasts that include and exclude the error correction term and it is clear that policy makers would have produced much better forecasts if the difference rule had been employed.[37] Figure 5B then shows

Table 4. Test for ARCH in Taylor Rule Residuals.

Sample	Test Statistic	Output Gap Measure
Full	11.64(0.00)	$\tilde{y}$
	14.15(0.00)	$\tilde{y}^{\mathrm{HP}}$
Pre-Greenspan	6.61(0.01)	$\tilde{y}$
	11.20(0.00)	$\tilde{y}^{\mathrm{HP}}$
Greenspan	0.69(0.41)	$\tilde{y}$
	3.26(0.07)	$\tilde{y}^{\mathrm{HP}}$
	2.82(0.09)	$\tilde{y}^{\mathrm{T3}}$
	0.12(0.73)	$\tilde{y}$
	1.36(0.24)	$\tilde{y} - \tilde{y}^{\mathrm{T}}$

Note: LM test statistic for the presence of ARCH (1) in the residuals for the Taylor rules defined in the last column. All Taylor rules use CPI inflation and the output gap proxy shown in the last column. *p*-values are given in parenthesis. Sample details are given in Table 1.

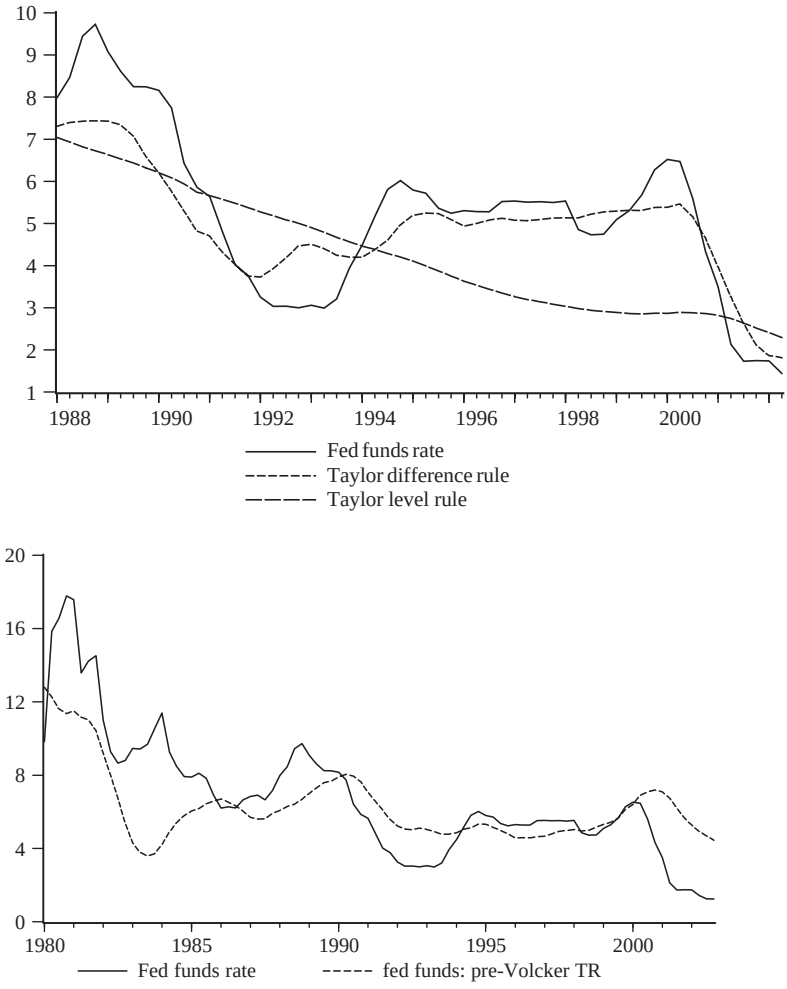


Fig. 5. (A) Implied Interest Rate During Greenspan's Tenure: Level vs. Difference Rules. (B) Counterfactual: The Volcker and Greenspan Tenures: Using a Pre-Volcker Taylor Rule. *Note:* The top figure shows the implied rules during the Greenspan era using a level versus a difference rule. See Eqs. (11) and (12). In Fig. (B), a Counterfactual is conducted wherein a level rule estimated during the Pre-Volcker era is used to generate the implied fed funds rate during the Greenspan sample.

what the Fed funds rate would have been if the coefficients in the Taylor rule estimated for the pre-Volcker era (1959.1–1979.4) had been used to predict interest rates during Greenspan's tenure. While policy would have clearly been too loose until about 1990, interest rates are fairly close to the actual Fed funds rate after that. On the face of it, one may be tempted to conclude that there is less to the choice of Fed chairman than priors might lead one to believe. However, if we compare the RMSE based on the implied interest rates we find, on balance, that there is a notable difference in monetary policy performance across various Fed chairmen (also see Romer & Romer, 2003).

6. SUMMARY AND LESSONS LEARNED

This paper has considered a variety of time series related questions that arise when estimating Taylor's rule. Focusing on the varied contributions of Clive Granger it was argued that most estimates have ignored the consequences of the unbalanced nature of Taylor's original specification. This means that some of the variables that make up Taylor's rule are either under-differenced or possibly over-differenced. Moreover, the extant literature has, with very few exceptions, omitted the possibility of cointegration among the variables in the Taylor rule. Augmenting Taylor's specification with an error correction term adds significantly to explaining interest rate movements, and consequently to our understanding of the conduct of monetary policy in the US since the late 1950s. Finally, there is also significant evidence of asymmetries in the conduct of monetary policy, at least during the Greenspan era. This highlights another of Clive Granger's remarkable contributions to econometric analysis, namely the importance of nonlinearities inherent in many economic relationships. A growing number of papers (e.g., Rabanal, 2004) are beginning to underscore this point.

NOTES

1. Goodhart (2004) casts doubt on the ability of expressions such as (1) to capture the preferences of central bankers if the regressions use the proper data. In particular, if a central bank has an inflation target and successfully meets it then the correlation in historical interest rate data and inflation data ought to vanish as central banks should always set the interest rate consistent with the target level of inflation.

2. Stability conditions are not necessarily as clear-cut as suggested above. See Woodford (2003, chapter 2).

3. They credit Mankiw and Shapiro (1985, 1986) for drawing attention to this problem. Also, see Enders (2004, pp. 102, 212), and Johnston and Dinardo (1997, p. 263).

4. For example, an augmented Dickey–Fuller test equation for the presence of a unit root is an unbalanced regression. Note, however, that standard inference on the $I(0)$ series in the test equation is not appropriate. Hence, one must rely on a non-standard distribution for hypothesis testing (e.g., see Hamilton, 1994, pp. 554–556).

5. Some might claim that non-stationarity in the interest rate makes no economic sense. Yet, empirical work relies on specifically chosen samples and, if there are sufficiently large shocks, the appearance of non-stationarity is difficult to reject.

6. Siklos (1999, 2002) also finds that inflation targeting central banks may indulge in relatively less interest rate smoothing behavior. Also, see Goodhart (2004).

7. English, Nelson, and Sack (2003) note that this observational equivalence problem may be overcome with a model estimated in first differences. Siklos (2004) also recognizes the near unit root nature of nominal interest rates and notes that a Taylor rule in first differences may also be more suitable for the analysis of countries that formally (or informally) target inflation. Using this type of formulation for the Taylor rule in a panel setting he finds that there are significant differences in the conduct of monetary policy between inflation and non-inflation targeting countries.

8. Indeed, since the real interest rate variable incorporates one or more possible cointegrating relationships, the absence of an error correction term in most Taylor rule equations is surprising. Further, given well-documented shifts in monetary policy, it is conceivable that a cointegrating relationship may be turned on, or off, in a regime-sensitive manner. It is precisely this type of consideration that led Siklos and Granger (1997) to propose regime-dependent cointegration.

9. Over-differencing can induce a non-invertible moving average component in the error term. The problem has been known for some time (e.g., Plosser & Schwert, 1977) and its emergence as an econometric issue stems from the spurious regression problem made famous by Granger and Newbold (1974).

10. For example, in the US, the first (or advance) release of real GDP data for each quarter is not available until about one month after the end of the quarter. The final release is not available until three months after the quarter ends. In addition, historical data are often revised. Hence, while Taylor rulesare usually estimated using final data, policy decisions often necessitate reliance on real-time data (see McCallum, 1998; Orphanides, 2001, 2003a, b; Orphanides & van Norden, 2002).

11. See also Smets (1998), Isard, Laxton, and Eliasson (1999), Orphanides et al. (1999), Muscattelli and Trecroci (2000) on Taylor rules and output gap and model uncertainty.

12. The basic idea is to generate a wide variety of estimates from different specifications and pool these to form combined estimates and confidence intervals. Indeed, Granger and Jeon (2004) apply their technique to the US data with the Taylor rule serving as the focal point of the technique's illustration.

13. Williams (2004) also points out that estimating a first differenced version of (1) reduces the problems associated with the uncertainty surrounding the measurement of the equilibrium real interest rate (i.e., the constant term in the Taylor rule).

14. Additionally, the inflation and output gap variables in (1) or (2) capture the trade-off between inflation and output variability and there are sound arguments, since at least Friedman (1977), to expect that an underlying long-run relationship exists between these two variables

15. In the event an interest rate smoothing parameter is present in the Taylor, the DW test statistic is inappropriate. One might instead resort to the use of the Durbin-h-test statistic.

16. Giannoni and Woodford (2003) consider the robustness of alternative policy rules and find in favor of a rule written in regression form as $i_t = \mu + \rho_1 i_{t-1} + \rho_2 \Delta i_{t-1} + \phi_\pi \pi_t + \phi_{\tilde{y}} \Delta \tilde{y}_t + \varepsilon_t$. The rule is very similar to the one specified by Judd and Rudebusch (1998) except that the Fed here reacts to the *change* in the output gap instead of its level. It is easy to show that this type of reaction function can be turned into a first difference version that includes an error correction term that reflects cointegration between the nominal interest rate and the output gap.

17. Space constraints prevent a review of Taylor rule estimates outside the US experience. Readers are asked to consult, for example, Clarida et al. (1998), Gerlach and Schnabel (2000), Gerlach-Kristen (2003), Gerdesmeier and Roffia (2004), and Siklos and Bohl (2005a, b).

18. It is far from clear that GMM provides different estimates of the parameters in a Taylor rule. One reason is that the literature has, until recently, almost completely ignored the issue of the choice of instruments and their relevance. Siklos and Bohl (2005a, b) show that this aspect of the debate may be quite crucial in deriving policy implications from Taylor rules.

19. In addition, the weights in Taylor-type rules also depend on how the rule interacts with other equations and variables in an economic structural model.

20. Clarida et al. (1998) perform some unit root tests and report mixed results in terms of the order of integration. They argue for the stationarity of the included variables based on the argument that the unit root tests have low power.

21. We intend to make available at http://www.wlu.ca/~wwwsbe/faculty/psiklos/home.htm the raw data and any programs used to generate the results. To facilitate access to the results we have relied on Eviews 5.0 for estimation.

22. The relevant data were downloaded from http://research.stlouisfed.org/fred2/.

23. Using the standard smoothing parameter of 1,600.

24. A simple estimate is based on the regression $\Delta\pi_t = \mu + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \gamma X_t + v_t$ where u is the unemployment rate, π is inflation, and X is a vector of other variables. We set $X = 0$ and used the CPI all items to obtain separate estimates for the 1957–1990 and 1991–2003 samples, which yielded estimates of 6.05% and 4.29%, respectively, where $NAIRU = -\mu/(\beta_1 + \beta_2)$.

25. Other than the NBER dates, the remaining data were kindly made available by Justin Wolfers and can also be downloaded from http://bpp.wharton.upenn.edu/jwolfers/personal_page/data.html.

26. Not surprisingly, the various core inflation measures are smoother than the ones that include all items and while, in general, the CPI and PCE based measures are similar there are some differences in the volatility of the series.

27. This view was not universally shared within the FOMC. Meyer (2004, p. 213), for example, did not believe that "…the Fed's effort to reduce inflation was somehow

responsible for the wave of innovation that propelled the acceleration of productivity in the second half of the 1990s.''

28. A highly readable account of the inspiration for the cointegration concept is contained in Hendry (2004).

29. The type of cointegration more or less implicit in Giannoni and Woodford (2003) is not generally found in the present data. Hence, we do not pursue this line of enquiry.

30. The test statistic for the null that the output gap does not belong in the cointegrating relationship is $\chi^2_2 = 5.80$ (p-value $= 0.05$). The test statistic for the null that i_t is weakly exogenous is $\chi^2_2 = 3.92$ (0.14). Both results are based on full sample estimates.

31. Another potential cointegrating relationship is the one between short and long rates (i.e., the term spread) but this variable proved to be statistically insignificant in almost all the specifications and samples considered.

32. Kozicki (1999, Table 1) reports an estimate of 2.34% when CPI inflation is used for the 1987–1997 sample. Our estimates of the equilibrium real rate (see Table 3) are sharply higher.

33. Where possible, estimates for the full sample were also generated but the results were just as disappointing as the ones discussed below. Estimating forecast-based rules pose a variety of difficulties. For example, Consensus and Economist forecasts used here are produced monthly and refer to forecasts for the subsequent calendar year. As the horizon until the forecast period narrows the information set available to forecasters expands and it is conceivable, if not likely, that the interest rate assumption used in the forecast has changed. No special adjustments were made here for this problem and it is not obvious how the correction ought to be made.

34. Interestingly, Kuttner (2004) reaches the same conclusion using an international data set. However, he does not consider the possibility that the averaging of forecasts might produce better results.

35. This conclusion is robust to whether we use actual output gap estimates versus forecasts of real GDP growth or the output gap (OECD) or to whether add an error correction term to the Taylor rule.

36. The results also lend further support for the findings first reported in Enders and Siklos (2001) who introduced the momentum model in tests for threshold cointegration (also see Enders & Granger, 1998).

37. The RMSE for the case where the error correction included is 0.95 and 1.92 when the error correction is excluded. Clements and Hendry (1995) discuss conditions under which incorporating an error correction term can improve forecasts.

REFERENCES

Ball, L. (1997). *Efficient rules for monetary policy*. NBER Working Paper no. 5952, March.

Banerjee, A., Dolado, J., Galbraith, J. W., & Hendry, D. F. (1993). *Co-Integration, error-correction, and the econometric analysis of non-stationary data*. Oxford: Oxford University Press.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly*, *20*, 451–468.

Bernanke, B. S. (2004a). "Fedspeak". Speech delivered at the annual meeting of the American economic association, San Diego, CA, January, available from www.federalreserve.gov

Bernanke, B. S. (2004b)."Gradualism." Remarks at an economics luncheon co-sponsored by the Federal Reserve Bank of San Francisco and the University of Washington, 20 May, available from www.federalreserve.gov

Bernanke, B. S., & Gertler, M. (1999). Monetary policy and asset price volatility. *Economic Review, Federal Reserve Bank of Kansas City (Fourth Quarter)*, 17–51.

Bernanke, B. S., & Mihov, I. (1998). Measuring monetary policy. *Quarterly Journal of Economics*, *113*(August), 869–902.

Boschen, J. F., & Mills, L. O. (1995). The relation between narrative and money market indicators of monetary policy. *Economic Inquiry*, *33*(January), 24–44.

Burdekin, R. C. K., & Siklos, P. L. (1999). Exchange regimes and shifts in inflation persistence: Does nothing else matter? *Journal of Money, Credit and Banking*, *31*(May), 235–247.

Castelnuovo, E. (2003). Taylor rules, omitted variables, and interest rate smoothing in the US. *Economics Letters*, *81*, 55–59.

Castelnuovo, E., & Surico, P. (2004). Model uncertainty, optimal monetary policy and the preferences of the Fed. *Scottish Journal of Political Economy*, *51*(February), 105–126.

Cecchetti, S., Genberg, H., Lipsky, J., & Wadhwani, S. (2000). Asset prices and central bank policy. *Geneva Reports on the World Economy* (Vol. 2). Geneva: International Center for Monetary and Banking Studies; London: Centre for Economic Policy Research.

Christensen, A. M., & Nielsen, H. B. (2003). *Has US monetary policy followed the Taylor rule? A cointegration analysis 1988–2002*. Working Paper, September 9.

Clarida, R., Gali, J., & Gertler, M. (1998). Monetary policy rules in practice some international evidence. *European Economic Review*, *42*(6), 1033–1067.

Clarida, R., Gali, J., & Gertler, M. (2000). Monetary policy and macroeconomic stability: Evidence and some theory. *The Quarterly Journal of Economics*, *115*(1), 147–180.

Clements, M. P., & Hendry, D. F. (1993). Forecasting in cointegrated systems. *Journal of Applied Econometrics*, *10*, 127–146.

Collins, S., & Siklos, P. (2004). Optimal monetary policy rules and inflation targets: Are Australia, Canada, and New Zealand different from the U.S.? *Open Economies Review*, *15*(October), 347–62.

Crowder, W., & Hoffman, D. (1996). The long-run relationship between nominal interest rates and inflation: The fisher equation revisited. *Journal of Money, Credit and Banking*, *28*, 102–118.

Culver, S. E., & Papell, D. H. (1997). Is there a unit root in the inflation rate? Evidence from sequential break and panel data models. *Journal of Applied Econometrics*, *12*, 435–444.

Dennis, R. (2003). *The policy preferences of the U.S. federal reserve*. Working Paper, Federal Reserve Bank of San Francsico, July.

Dolado, J., Pedrero, R., & Ruge-Murcia, F. J. (2004). Nonlinear policy rules: Some new evidence for the U.S. *Studies in Nonlinear Dynamics and Econometrics*, *8*(3), 1–32.

Domenech, R., Ledo, M., & Tanguas, D. (2002). Some new results on interest rate rules in EMU and in the U.S. *Journal of Economics and Business*, *54*, 431–446.

Dueker, M., & Rasche, R. H. (2004). Discrete policy changes and empirical models of the federal funds rate. *Review of the Federal Reserve Bank of St. Louis* (forthcoming).

Elliott, G., & Stock, J. H. (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory*, *10*, 672–700.
Enders, W. (2004). *Applied econometric time series* (2nd ed.). New York: Wiley.
Enders, W., & Granger, C. W. J. (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.
Enders W., & Siklos, P. L. (2001). Cointegration and threshold adjustment. *Journal of Business and Economic Statistics*, *19*(April), 166–177.
English, W. B., Nelson, W. R., & Sack, B. (2003). Interpreting the significance of the lagged interest rate in estimated monetary policy rules. *Contributions to Macroeconomics*, *3*(1), Article 5.
Favero, C. A., & Rovelli, R. (2003). Macroeconomic stability and the preferences of the fed: A formal analysis, 1961–98. *Journal of Money, Credit and Banking*, *35*(August), 545–556.
Friedman, M. (1977). Nobel lecture: Inflation and unemployment. *Journal of Political Economy*, *85*(June), 451–72.
Fuhrer, J., & Tootell, G. (2004). *Eyes on the prize: How did the fed respond to the stock market?* Public Policy Discussion Papers 04-2, Federal Reserve Bank of Boston, June.
Gerdesmeier, D., & Roffia, B. (2004). Empirical estimates of reaction functions for the Euro area. *Swiss Journal of Economics and Statistics*, *140*(March), 37–66.
Gerlach-Kristen, P. (2003). *A Taylor rule for the Euro area*. Working Paper, University of Basel.
Gerlach, S., & Schnabel, G. (2000). The Taylor rule and interest rates in the EMU area. *Economics Letters*, *67*, 165–171.
Giannoni, M. P., & Woodford, M. (2003). How forward looking is monetary policy? *Journal of Money, Credit and Banking*, *35*(December), 1425–1469.
Goodfriend, M. (1991). Interest rates and the conduct of monetary policy. *Carnegie-Rochester Conference Series on Public Policy*, *34*, 7–30.
Goodhart, C. A. E. (1999). Central bankers and uncertainty. *Bank of England Quarterly Bulletin*, *39*(February), 102–114.
Goodhart, C. A. E. (2004). *The monetary policy committee's reaction function: An exercise in estimation*. Working Paper, Financial Markets Group, London School of Economics.
Granger, C. W. J. (2004). Time series analysis, cointegration, and applications. *American Economic Review*, *94*(June), 421–425.
Granger, C. W. J., & Jeon, Y. (2004). Thick modelling. *Economic Modelling*, *21*, 323–343.
Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*, 111–120.
Hamalainen, N. (2004). *A survey of Taylor-type monetary policy rules*. Working Paper, Department of Finance, Ministry of Finance, Canada.
Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.
Hendry, D. F. (1995). *Dynamic econometrics*. Oxford: Oxford University Press.
Hendry, D. F. (2004). The nobel memorial price for Clive W. J. Granger. *Scandinavian Journal of Economics*, *106*(2), 187–213.
Hetzel, R. L. (2000). The Taylor rule: Is it a useful guide to understanding monetary policy? *Federal Reserve Bank of Richmond Economic Quarterly*, *86*, 1–33.
Isard, P., Laxton, D., & Eliasson, A.-C. (1999). Simple monetary policy rules under model uncertainty. *International Tax and Public Finance*, *6*, 537–577.
Johnston, J., & Dinardo, J. (1997). *Econometric methods* (4th ed.). New York: McGraw-Hill.

Judd, J. P., & Rudebusch, G. D. (1998). Taylor's rule and the Fed: 1970–1997. *Federal Reserve Bank of San Francisco Economic Review*, *3*, 3–16.

Kozicki, S. (1999). How useful are Taylor rules for monetary policy? *Federal Reserve Bank of Kansas City Economic Review*, *84*(2), 5–33.

Kuttner, K. N. (2004). A snapshot of inflation targeting in its adolescence. In: K. Christopher & S. Guttman (Eds), *The future of inflation targeting* (pp. 6–42). Sydney, Australia: Reserve Bank of Australia.

Lanne, M. (1999). Near unit roots and the predictive power of yield spreads for changes in long-term interest rates. *Review of Economics and Statistics*, *81*, 393–398.

Lanne, M. (2000). Near unit roots, cointegration, and the term structure of interests rates. *Journal of Applied Econometrics*, *15*(September–October), 513–529.

Laubach, T., & Williams, J. C. (2003). Measuring the natural rate of interest. *Review of Economics and Statistics*, *85*(November), 1063–1070.

Leitemo, K., & Røisland, Ø. (1999). Choosing a monetary policy regime: Effects on the traded and non-traded sectors. University of Oslo and Norges Bank.

Levin, A., Wieland, V., & Williams, J. C. (1999). Robustness of simple policy rules under model uncertainty. In: J. B. Taylor (Ed.), *Monetary policy rules*. Chicago: University of Chicago Press.

MacKinnon, J. G., Haug, A., & Michelis, L. (1999). Numerical distribution functions of likelihood ratio tests for cointegration. *Journal of Applied Econometrics*, *14*, 563–577.

Mankiw, N. G., & Shapiro, M. D. (1985). Trends, random walks, and tests of the permanent income hypothesis. *Journal of Monetary Economics*, *16*(September), 165–174.

Mankiw, N. G., & Shapiro, M. D. (1986). Do we reject too often? Small sample properties of tests of rational expectations models. *Economics Letters*, *20*, 139–145.

McCallum, B. T. (1998). Issues in the design of monetary policy rules. In: J. B. Taylor & M. Woodford (Eds), *Handbook of macroeconomics*. Elsevier.

McCallum, B. T. (2001). Should monetary policy respond strongly to output gaps? *American Economic Review*, *91*(2), 258–262.

McCallum, B. T., & Nelson, E. (1999). Performance of operational policy rules in an estimated semiclassical structural model. In: J. B. Taylor (Ed.), *Monetary policy rules*. Chicago: University of Chicago Press.

Medina, J. P., & Valdés, R. (1999). Optimal monetary policy rules when the current account matters. Central Bank of Chile.

Meyer, L. (2004). *A Term at the Fed*. New York: HarperCollins.

Muscatelli, V., Tirelli, A. P., & Trecroci, C. (1999). *Does institutional change really matter? Inflation targets, central bank reform and interest rate policy in the OECD countries*. CESifo Working Paper No. 278, Munich.

Muscattelli, A., & Trecroci, C. (2000). Monetary policy rules, policy preferences, and uncertainty: Recent empirical evidence. *Journal of Economic Surveys*, *14*(5), 597–627.

Ng, S., & Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, *69*(November), 1519–1554.

Oliner, S. D., & Sichel, D. E. (2000). The resurgence of growth in the late 1990s: Is information technology the story? *Journal of Economic Perspective*, *14*(Fall), 3–23.

Orphanides, A. (2001). Monetary policy rules based on real-time data. *American Economic Review*, *91*(4), 964–985.

Orphanides, A. (2003a). Monetary policy evaluation with noisy information. *Journal of Monetary Economics*, *50*(3), 605–631.

Orphanides, A. (2003b). Historical monetary policy analysis and the Taylor rule. *Journal of Monetary Economics*, *50*(5), 983–1022.

Orphanides, A., & van Norden, S. (2002). The unreliability of output gap estimates in real time. *Review of Economics and Statistics*, *84*(4), 569–583.

Orphanides, A., Porter, R. D., Reifschneider, D., Tetlow, R., & Finan, F. (1999). *Errors in the Measurement of the output gap and the design of monetary policy*. Finance and Economics Discussion Series, 1999–45, Federal Reserve Board, August.

Osterholm, P. (2003). *The Taylor rule: A spurious regression*. Working Paper, Uppsala University.

Ozlale, U. (2003). Price stability vs. output stability: Tales from three federal reserve administrations. *Journal of Economic Dynamics and Control*, 9(July), 1595–1610.

Phillips, P. C. B. (1986). Understanding spurious regression in econometrics. *Journal of Econometrics*, *33*, 311–340.

Phillips, P. C. B. (1989). Regression theory for nearly-integrated time series. *Econometrica*, *56*, 1021–1043.

Plosser, C. I., & Schwert, G. W. (1977). Estimation of a non-invertible moving average process: The case of over-differencing. *Journal of Econometrics*, *6*(September), 199–224.

Rabanal, P. (2004). *Monetary policy rules and the U.S. business cycle: Evidence and implications*. International Monetary Fund Working Paper WP/04/164, September.

Rapach, D. E., & Wohar, M. E. (2005). Regime changes in international real interest rates: Are they a monetary phenomenon? *Journal of Money, Credit and Banking*, *37*(October), 887–906.

Romer, C. D., & Romer, D. H. (1989). Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. *NBER Macroeconomics Annual 1989* (pp. 121–170). Cambridge, MA: The MIT Press.

Romer, C. D., & Romer, D. H. (2003). *Choosing the federal reserve chair: Lessons from history*. Working Paper, University of California, Berkeley.

Rudebusch, G. D. (2001). Is the fed too timid? Monetary policy in an uncertain world. *Review of Economics and Statistics*, *83*(2), 203–217.

Rudebusch, G. D. (2002). Term structure evidence on interest rate smoothing and monetary policy inertia. *Journal of Monetary Economics*, *49*, 1161–1187.

Rudebusch, G. D., & Svensson, L. E. O. (1999). Policy rules for inflation targeting. In: J. B. Taylor (Ed.), *Monetary policy rules*. Chicago: University of Chicago Press.

Ruth, K. (2004). *Interest rate reaction functions for the Euro area: Evidence from panel data analysis*. Deutsche Bundesbank Working Paper 33/2004.

Sack, B. (1998). Does the fed act gradually? A VAR analysis. *Journal of Monetary Economics*, *46*(August), 229–256.

Sack, B., & Wieland, V. (2000). Interest-rate smoothing and optimal monetary policy: A review of recent empirical evidence. *Journal of Economics and Business*, *52*, 205–228.

Siklos, P. L. (1999). Inflation target design: Changing inflation performance and persistence in industrial countries. *Review of the Federal Reserve Bank of St. Louis*, *81*(March/April), 47–58.

Siklos, P. L. (2002). *The changing face of central banking: Evolutionary trends since World War II*. Cambridge: Cambridge University Press.

Siklos, P. L. (2004). Central bank behavior, the institutional framework, and policy regimes: Inflation versus noninflation targeting countries. *Contemporary Economic Policy*, (July), 331–343.

Siklos, P. L., & Bohl, M. T. (2005a). *The role of asset prices in Euro area monetary policy: Specification and estimation of policy rules and implications for the ECB*. Working Paper, Wilfrid Laurier University.

Siklos, P. L., & Bohl, M. T. (2005b). *Are inflation targeting central banks relatively more forward looking*? Working Paper, Wilfrid Laurier University.

Siklos, P. L., & Granger, C. W. J. (1997). Regime-sensitive cointegration with an illustration from interest rate parity. *Macroeconomic Dynamics*, *1*(3), 640–657.

Sims, C., Stock, J. H., & Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, *58*(January), 113–144.

Smets, F. (1998). *Output gap uncertainty: Does it matter for the Taylor rule*? BIS Working Papers, no. 60, November.

Söderlind, P., Söderström, U., & Vredin, A. (2003). Taylor rules and the predictability of interest rates. Working Paper no. 147, Sveriges Riksbank.

Staiger, D., Stock, J. H., & Watson, M. W. (1997). The NAIRU, unemployment and monetary policy. *Journal of Economic Perspectives*, *11*(Summer), 33–50.

Svensson, L.E.O. (2003). What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules. *Journal of Economic Literature*, *XLI*(June), 426–477.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie–Rochester Conference Series on Public Policy*, *39*, 195–214.

Taylor, J. B. (1999). A historical analysis of monetary policy rules. In: J. B. Taylor (Ed.), *Monetary policy rules*. Chicago: University of Chicago Press.

Tchaidze, R. (2004). *The greenbook and U.D. monetary policy*. IMF Working Paper 04/213, November.

Walsh, C. (2003a). Speed limit policies: The output gap and optimal monetary policy. *American Economic Review*, *93*(March), 265–278.

Walsh, C. (2003b). Implications for a changing economic structure for the strategy of monetary policy. In: Monetary Policy and Uncertainty: Adapting to a Changing Economy, A symposium sponsored by the Federal Reserve Bank of Kansas City Jackson Hole, Wyoming, August 28–30.

Williams, J. C. (2004). *Robust estimation and monetary policy with unobserved sructural change*. Federal Reserve Bank of San Francisco Working Paper 04–11, July.

Woodford, M. (2001). The Taylor rule and optimal monetary policy. *American Economic Review*, *91*(2), 232–237.

Woodford, M. (2003). *Interest and Prices*. Princeton: Princeton University Press.

BAYESIAN INFERENCE ON MIXTURE-OF-EXPERTS FOR ESTIMATION OF STOCHASTIC VOLATILITY

Alejandro Villagran and Gabriel Huerta

ABSTRACT

The problem of model mixing in time series, for which the interest lies in the estimation of stochastic volatility, is addressed using the approach known as Mixture-of-Experts (ME). Specifically, this work proposes a ME model where the experts are defined through ARCH, GARCH and EGARCH structures. Estimates of the predictive distribution of volatilities are obtained using a full Bayesian approach. The methodology is illustrated with an analysis of a section of US dollar/German mark exchange rates and a study of the Mexican stock market index using the Dow Jones Industrial index as a covariate.

1. INTRODUCTION

In options trading and in foreign exchange rate markets, the estimation of volatility plays an important role in monitoring radical changes over time of

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 277–296
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20030-0

key financial indexes. From a statistical standpoint, volatility refers to the variance of the underlying asset or return, conditional to all the previous information available until a specific time point.

It is well known that the volatility of a financial series tends to change over time and there are different types of models to estimate it: continuous-time or discrete-time processes. This paper discusses a model that falls in the second category. Discrete-time models are divided into those for which the volatility is ruled by a deterministic equation and those where the volatility has a stochastic behavior. Among the former, we have the autoregressive conditional heteroskedastic (ARCH) model introduced by Engle (1982) which provided a breakthrough in the modeling and estimation of time-varying conditional variance. Extensions to this seminal model are the generalized ARCH (GARCH) of Bollerslev (1986), the exponential GARCH (EGARCH) of Nelson (1991), the Integrated GARCH (IGARCH) and the GARCH in mean (GARCH-M). Additionally, models like the stochastic volatility (SV) of Melino and Turnbull (1990) give a discrete-time approximation for a continuous diffusion processes used in option pricing.

These variety of models gives a portfolio of options to represent volatility, but no agreement to decide which is the best approach. Given this difficulty, a new source of modeling arised: *mixture-of-models.* Since the only agreement seems to be that no real process can be completely explained by one model, the idea of model mixing, to combine different approaches into a unique representation, is very interesting. There are many types of methods of mixture models to estimate volatility. For instance, Wong and Li (2001) proposed a mixture of ARCH models with an autoregressive component to model the mean (MAR-ARCH) and for which they use the Expectation Maximization or EM algorithm to produce point estimation of the volatility. Another approach is given by Tsay (2002), who considered a mixture of ARCH and GARCH models by Markov Switching. Vrontos, Dellaportas, and Politis (2000) used *reversible jump* Markov chain Monte Carlo (MCMC) methods to predict a future volatility via model averaging of GARCH and EGARCH models. Both of these papers only considered a mixture of two models. Furthermore, Huerta, Jiang, and Tanner (2003) discussed the neural networks approach known as Hierarchical Mixture-of-Experts (HME), which is a very general and flexible approach of model mixing since it incorporates additional exogenous information, in the form of covariates or simply time, through the weights of the mixture. The example shown in that paper makes comparisons between a *difference-stationary* and

a *trend-stationary* model using time as the only covariate. Additionally, Huerta, Jiang, and Tanner (2001), considers an HME model including AR, ARCH and EGARCH models and obtained point estimates of volatility using the EM algorithm. However, that paper does not report interval estimates via a full Bayesian approach.

In this paper, we use *Mixture-of-Experts* (ME), which is a particular case of HME, to build a mixture of ARCH, GARCH and EGARCH models. Through a full Bayesian approach based on MCMC methods, we show how to estimate the posterior distribution of the parameters and the posterior predictive distribution of the volatility, which is a very complicated function of the mixture representation. Additionally, we show how to obtain point estimates of the volatility, but also the usually unavailable forecast intervals. The paper proceeds in the following way. In Section 2, we offer an introduction to ME and HME in the context of time series modeling. In Section 3, we give the details of our volatility ME model along with the MCMC specifications to implement a full Bayesian approach to the problem of estimating volatility. Section 4 illustrates our methodology in the context of two financial applications and Section 5 provides some conclusions and extensions.

2. MIXTURE MODELING

Hierarchical Mixture of Experts (HME) was first introduced in the seminal paper by Jordan and Jacobs (1994) and it is based on mixing models to construct a *neural network* the using logistic distribution. This approach allows for model comparisons and a representation of the mixture weights as a function of time or other covariates. Additionally, the elements of the mixture, also known as experts, are not restricted to a particular parametric family, which allows for very general model comparisons.

The model considers a response time series $\{y_t\}$ and a time series of covariates or exogenous variables $\{x_t\}$. Let $f_t(y_t|\mathscr{F}_{t-1},\chi;\theta)$ be the probability density function (pdf) of y_t conditional on θ, a vector of parameters, χ the σ-algebra generated by the exogenous information $\{x_t\}_0^n$, and for each t, $\mathscr{F}_{t-1}$ is the σ-algebra generated by $\{y_s\}_0^{t-1}$, the previous history about the response variable up to time $t-1$. Usually, it is assumed that this conditional pdf only depends on χ through x_t.

In the methodology of HME the pdf f_t is assumed to be a mixture of conditional pdfs of simpler models (Peng, Jacobs, & Tanner, 1996). In the

context of time series, the mixture could be represented by a finite sum

$$f_t(y_t|\mathscr{F}_{t-1},\chi;\theta)=\sum_J g_t(J|\mathscr{F}_{t-1},\chi;\gamma)\pi_t(y_t|\mathscr{F}_{t-1},\chi,J;\eta)$$

where the functions $g_t(\cdot\,|\,\cdot,\cdot;\gamma)$ are the mixtures weights; $\pi_t(\cdot\,|\,\cdot,\cdot,J;\eta)$ are the pdfs of simpler models defined by the label J; γ and η are sub-vectors of the parameter vector θ.

The models that are being mixed in HME are commonly denoted as *experts.* For example, in time series, one expert could be an AR(1) model, another expert could be a GARCH(2,2) model. Also, the experts could be models that belong to the same class but with different orders or number of parameters. For example, all the experts are AR model but with different orders and different values of the *lag* coefficients. The extra hierarchy in HME partitions the space of covariates into O "overlays". In each overlay we have M competing models so that the most appropriate model will be assigned a higher weight.

For this hierarchical mixture, the expert index J, could be expressed as $J=(o,m)$, where the overlay index o takes a value in the set $\{1,\ldots,O\}$ and the model type index m takes a value in $\{1,\ldots,M\}$. The mixture model can be rewritten as

$$f_t(y_t|\mathscr{F}_{t-1},\chi;\theta)\sum_{o=1}^{O}\sum_{m=1}^{M} g_t(o,\ m|\mathscr{F}_{t-1},\chi;\gamma)\pi_t(y_t|\mathscr{F}_{t-1},\chi,o,\ m;\eta)$$

Within the neural network terminology, the mixture weights are known as *gating functions*. In difference to other approaches these weights have a particular parametric form that may depend on the previous history, exogenous information or exclusively on time. This makes the weights evolve across time in a very flexible way.

Specifically, it is proposed that the mixture weights have the form,

$$g_t(o,\ m|\mathscr{F}_{t-1},\chi;\gamma)=\left\{\frac{e^{v_o+u_o^T W_t}}{\sum_{s=1}^{O} e^{v_s+u_s^T W_t}}\right\}\left\{\frac{e^{v_{m|o}+u_{m|o}^T W_t}}{\sum_{l=1}^{M} e^{v_{l|o}+u_{l|o}^T W_t}}\right\}$$

where the v's and u's are parameters, which are components of γ; W_t an input at time t, which is measurable with respect to the σ-algebra induced by $\mathscr{F}_{t-1}\cup\chi$. In this case, γ includes the following components: $v_1,u_1,\ldots,v_{O-1},u_{O-1},v_{1|1},u_{1|1},\ldots,v_{M-1|1},u_{M-1|1},\ldots,v_{M-1|O},u_{M-1|O}$. For identifiability of the mixture weights, we set $v_O=u_O=v_{M|o}=u_{M|o}=0$ for all $o=1,\ldots,O$. This restriction guarantees that the gating functions are uniquely identified by γ as shown in Huerta et al. (2003). Both terms that

define the mixture weights follow a multinomial logistic pdf where the first term describes the probability of a given overlay and the second term, the probability of a model within overlay. Each of these probabilities being a function of the input W_t. Mixtures of time series models for estimating volatility has also been considered in Wong and Li (2001). However, these authors only look at the problem from a point estimation perspective.

Inferences on the parameter vector θ can be based in the log-likelihood function

$$\mathscr{L}_n(\cdot) = \frac{1}{n}\sum_{t=1}^{n} \log f_t(y_t|\mathscr{F}_{t-1}, \chi; \cdot)$$

To obtain the maximum likelihood estimator of θ, $\hat{\theta} = \arg\max \mathscr{L}_n(\cdot)$, it is possible to use the EM algorithm as described in Huerta et al. (2003). A general presentation of the EM algorithm appears in Tanner (1996).

After the MLE, $\hat{\theta}$, is obtained, the interest focuses in the evaluation of the weights assigned to each of the M models as a function of time t. Primarily, there are two ways to achieve this, the first one is via the conditional probability of each model m defined by

$$P_t(m|y_t,\ \mathscr{F}_{t-1},\ \chi,\ \theta) \equiv h_m(t) \equiv \sum_{o=1}^{O} h_{om}(t;\theta)$$

where conditional refers to the actual observation at time t, y_t.

The second approach is to consider the unconditional probability at time t given by

$$P_t(m|\mathscr{F}_{t-1},\ \chi,\ \theta) \equiv g_m(t) \equiv \sum_{o=1}^{O} g_{om}(t;\theta)$$

Point estimation of each probability can be obtained by evaluation at $\hat{\theta}$ or by computing the expected value with respect to the posterior distribution $\pi(\theta|\mathscr{F}_n,\ \chi)$.

The particular case of HME that we consider in this paper is $O = 1$, which is known as Mixture of Experts (ME). In the ME modeling it is assumed that the process that generates the response variable can be decomposed into a set of subprocesses defined over specific regions of the space of covariates. For each value of the covariate x_t, a 'label' r is chosen with probability $g_t(r|\chi, \mathscr{F}_{t-1}, \gamma)$. Given this value of r, the response y_t is generated from the conditional pdf $\pi_t(y_t|r, \chi, \mathscr{F}_{t-1}, \eta)$. The pdf of y_t conditional on the

parameters, the covariate and the response history is given by

$$f(y_t|\chi, \mathscr{F}_{t-1}, \theta) = \sum_{r=1}^{M} g_t(r|\chi, \mathscr{F}_{t-1}, \gamma)\pi_t(y_t|r, \chi, \mathscr{F}_{t-1}, \eta)$$

and the likelihood function is

$$L(\theta|\chi) = \prod_{t=1}^{n}\sum_{r=1}^{M} g_t(r|\chi, \mathscr{F}_{t-1}, \gamma)\pi_t(y_t|r, \chi, \mathscr{F}_{t-1}, \eta)$$

As with HME, the mixture probabilities associated to each value of r are defined through a logistic function

$$g_t(r|\chi, \mathscr{F}_{t-1}, \eta) \equiv g_r^{(t)} = \frac{e^{\xi_r}}{\sum_{h=1}^{M} e^{\xi_h}}$$

where $\xi_r = v_r + u_r^T W_t$, W_t is an input that could be a function of time, history and $\{x_t\}$. For identifiability, ξ_M is set equal to zero.

Inference about the parameters in the ME is simplified by augmenting the data with nonobservable indicator variables, which determine the type of model expert. For each time t, $z_r^{(t)}$ is a binary variable such that $z_r^{(t)} = 1$ with probability

$$h_r^{(t)} = \frac{g_r^{(t)}\pi_t(y_t|r, \chi, \mathscr{F}_{t-1}, \eta)}{\sum_{r=1}^{M} g_r^{(t)}\pi_t(y_t|r, \chi, \mathscr{F}_{t-1}, \eta)}$$

If $\chi' = \{(x_t, z^{(t)})\}_{t=1}^{n}$, where $z^{(t)}$ is the vector that includes all the indicator variables, the *augmented likelihood* for the ME model is

$$L(\theta|\chi') = \prod_{t=1}^{n}\prod_{r=1}^{M}\{g_r^{(t)}\pi_t(y_t|r, \chi, \mathscr{F}_{t-1}, \gamma)\}^{z_r^{(t)}}$$

In the following section, we discuss how to estimate a ME model by a Bayesian approach and with the experts being ARCH, GARCH and EGARCH models. We picked these experts as our model building blocks since these are the main conditional heteroskedasticity models used in practice as pointed out by the seminal papers of Engle (1982, 1995), Bollerslev (1986) and Nelson (1991). Also, Tsay (2002) and Vrontos, et al. (2000) mention that these models are interesting in practice due to their parsimony.

3. BAYESIAN INFERENCE ON MIXTURES FOR VOLATILITY

If the Bayesian paradigm is adopted, the inferences about θ are based on the posterior distribution $\pi(\theta|\underline{y})$. Bayes Theorem establishes that

$$\pi(\theta|\underline{y}) = \frac{f(\underline{y}|\theta)\pi(\theta)}{\int_\Theta f(\underline{y}|\theta)\,\mathrm{d}F^\pi(\theta)}$$

which defines the way to obtain the posterior distribution of θ through the prior $\pi(\theta)$ and the likelihood function $f(\underline{y}|\theta)$. However, for a ME or HME approach the marginal distribution of $\underline{y}$, $\int_\Theta f(\underline{y}|\theta)\,\mathrm{d}F^\pi(\theta)$ cannot be obtained analytically. We overcome this difficulty by using MCMC methods to simulate samples from $\pi(\theta|\underline{y})$. For more details about MCMC methods see Tanner (1996).

First, we assume that the prior distribution for $\theta = (\eta, \gamma)$ has the form

$$\pi(\theta) = \pi(\eta)\pi(\gamma)$$

so the expert parameter η and the weights or *gating* parameters γ are a priori independent. We define $\mathbf{Z} = \{\mathbf{z^{(t)}}; t = 1, \ldots, n\}$ and for each t, $\mathbf{z^{(t)}} = \{z_r^{(\mathbf{t})}; r = 1, \ldots, M\}$ is the set of indicator variables at t. Conditional on θ, $P(\mathbf{z}^{(t)}|\theta, \chi)$ is a Multinomial distribution with total count equal to 1 and cell probabilities $g_r(t,\gamma)$.

Our MCMC scheme is based on the fact that it is easier to obtain samples from the *augmented* posterior distribution $\pi(\theta, \mathbf{Z}|\mathscr{F}_n, \chi)$, instead of directly simulating values from $\pi(\theta|\mathscr{F}_n, \chi)$. This data augmentation principle was introduced by Tanner and Wong (1987). The MCMC scheme follows a Gibbs sampling format for which we iteratively simulate from the conditional distributions $\pi(\theta|\mathbf{Z}, \mathscr{F}_n, \chi)$ and $\pi(\mathbf{Z}|\theta, \mathscr{F}_n, \chi)$.

The conditional posterior $\pi(\mathbf{Z}|\theta, \mathscr{F}_n, \chi)$ is sampled through the marginal conditional posteriors $\pi(\mathbf{z}^{(t)}|\theta, \mathscr{F}_n, \chi)$ defined for each value of t. Given θ, $\mathscr{F}_n$ and χ, it can be shown that the vector $\mathbf{z}^{(t)}$ has a Multinomial distribution with total count equal to 1 and for which

$$P(z_r^{(t)} = 1|\theta, \mathscr{F}_n, \chi) = h_r(t;\theta) = \frac{g_r^{(t)}\pi_t(y_t|r, \mathscr{F}_{t-1}, \chi; \eta)}{\sum_{r=1}^M g_r^{(t)}\pi_t(y_t|r, \mathscr{F}_{t-1}, \chi; \eta)}$$

The vector $\theta = (\eta, \gamma)$ is sampled in two stages. First, η is simulated from the conditional posterior distribution $\pi(\eta|\gamma, \mathbf{Z}, \mathscr{F}_n, \chi)$ and then γ is sampled from the conditional posterior $\pi(\gamma|\eta, \mathbf{Z}, \mathscr{F}_n, \chi)$. By Bayes Theorem,

$$\pi(\eta|\gamma, \mathbf{Z}, \mathscr{F}_n, \chi) \propto \prod_{t=1}^n \prod_{r=1}^M f_t(y_t|\mathscr{F}_{t-1}, \chi, r; \eta)^{z_r^{(t)}}\pi(\eta)$$

Analogously,

$$\pi(\gamma|\eta, \mathbf{Z}, \mathscr{F}_n, \chi) \propto \prod_{t=1}^{n}\prod_{r=1}^{M} g_t(r|\mathscr{F}_{t-1}, \chi; \gamma)^{z_r^{(t)}}\pi(\gamma)$$

If η can be decomposed into a sub-collection of parameters η_r that are assumed a priori independent, the simulation for the full conditional for η is reduced to individual simulation of each η_r. If η_r is assigned a conjugate prior with respect to the pdf of the "r" expert, the simulation of η is straightforward. For γ, it is necessary to implement Metropolis–Hastings steps to obtain a sample from its full conditional distribution.

The specific details for the MCMC implementation depends on the type of expert models and prior distributions on model parameters. For example, Huerta et al. (2003) discussed a HME model with a full Bayesian approach where the experts are a 'difference-stationary' and a 'trend-stationary' model. The priors used on the parameters of their HME model were non-informative. Here, we consider the case of experts that allow volatility modeling.

It is well known that the volatility of a financial time series can be represented by ARCH, GARCH and EGARCH models. The properties of these models make them attractive to obtain forecasts in financial applications. We propose a ME that combines the models AR(1)-ARCH(2), AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1). Although the order of the autoregressions for the observations and volatility is low for these models, in practice it is usually not necessary to consider higher order models.

The elements of our ME model are, the time series of returns $\{y_t\}_1^n$, the series of covariates $\{x_t\}_1^n$, which in one of our applications presented in the next section it is simply time and in the other, it is the Dow Jones index (DJI). In any case, $\xi_r = \upsilon_r + u_r^T W_t$, where W_t is an input that depends on the covariates.

Our expert models will be parameterized in the following way:

AR(1)-ARCH(2)

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_{1,t} \quad \varepsilon_{1,t} \sim N(0, \sigma_{1,t}^2) \\ \sigma_{1,t}^2 &= \omega_1 + \alpha_{11}\varepsilon_{1,t-1}^2 + \alpha_{12}\varepsilon_{1,t-2}^2 \end{aligned}$$

AR(1)-GARCH(1,1)

$$\begin{aligned} y_t &= \phi_2 y_{t-1} + \varepsilon_{2,t} \quad \varepsilon_{2,t} \sim N(0, \sigma_{2,t}^2) \\ \sigma_{2,t}^2 &= \omega_2 + \alpha_{21}\varepsilon_{2,t-1}^2 + \alpha_{22}\sigma_{2,t-2}^2 \end{aligned}$$

AR(1)-EGARCH(1,1)

$$y_t = \phi_3 y_{t-1} + \varepsilon_{3,t} \quad \varepsilon_{3,t} \sim N(0, \sigma^2_{3,t})$$
$$\ln(\sigma^2_{3,t}) = \omega_3 + \alpha_{31} \ln(\sigma^2_{3,t-1}) + \alpha_{32}\varepsilon_{3,t-1} + \alpha_{33}(|\varepsilon_{3,t-1}| - E(|\varepsilon_{3,t-1}|))$$

For each expert $m = 1, 2, 3$, the ME will be represented by the following pdfs and gating functions:

Expert 1

$$g_t(1|\mathscr{F}_{t-1}, \chi; \gamma) = \frac{\exp\{\xi_1\}}{\sum_{r=1}^{3}\exp\{\xi_r\}} = \frac{\exp\{v_1 + u_1 W_t\}}{\sum_{r=1}^{3}\exp\{v_r + u_r W_t\}}$$
$$\pi_t(y_t|1, \mathscr{F}_{t-1}, \chi; \eta_1) = \frac{1}{\sqrt{2\pi\sigma^2_{1,t}}}\exp\left\{-\frac{1}{2\sigma^2_{1,t}}(y_t - \phi_1 y_{t-1})^2\right\}$$

Expert 2

$$g_t(2|\mathscr{F}_{t-1}, \chi; \gamma) = \frac{\exp\{\xi_2\}}{\sum_{r=1}^{3}\exp\{\xi_r\}} = \frac{\exp\{v_2 + u_2 W_t\}}{\sum_{r=1}^{3}\exp\{v_r + u_r W_t\}}$$
$$\pi_t(y_t|2, \mathscr{F}_{t-1}, \chi; \eta_2) = \frac{1}{\sqrt{2\pi\sigma^2_{2,t}}}\exp\left\{-\frac{1}{2\sigma^2_{2,t}}(y_t - \phi_2 y_{t-1})^2\right\}$$

Expert 3

$$g_t(3|\mathscr{F}_{t-1}, \chi; \gamma) = \frac{\exp\{\xi_3\}}{\sum_{r=1}^{3}\exp\{\xi_r\}} = \frac{1}{\sum_{r=1}^{3}\exp\{v_r + u_r W_t\}}$$
$$\pi_t(y_t|3, \mathscr{F}_{t-1}, \chi; \eta_3) = \frac{1}{\sqrt{2\pi\sigma^2_{3,t}}}\exp\left\{-\frac{1}{2\sigma^2_{3,t}}(y_t - \phi_3 y_{t-1})^2\right\}$$

As mentioned before, the vector $\theta = (\eta, \gamma)$ can be decomposed into two subsets, one that includes the expert parameters, $\eta = (\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \omega_1, \omega_2, \omega_3, \varphi_1, \varphi_2, \varphi_3)$ and another subvector that includes the gating function parameters, $\gamma = (u_1, u_2, v_1, v_2)$.

The likelihood function of the model is expressed as,

$$\mathscr{L}(\theta|\chi) = \prod_{t=1}^{n}\sum_{r=1}^{3} g_r^{(t)}\exp\{-\frac{1}{2\sigma^2_{r,t}}(y_t - \phi_r y_{t-1})^2\}$$

and the augmented likelihood function is,

$$\mathscr{L}(\theta|\chi') = \prod_{t=1}^{n}\prod_{r=1}^{3}\left\{g_r^{(t)}\exp\{-\frac{1}{2\sigma^2_{r,t}}(y_t - \phi_r y_{t-1})^2\}\right\}^{z_r^{(t)}}$$

The MCMC scheme to obtain posterior samples for this ME is based on the principle of "divide and conquer". For *Expert 1*, we assume that $\eta_1 = (\phi_1, \omega_1, \alpha_{11}, \alpha_{12})$ has prior distribution with components that are a priori

independent, $\pi(\eta_1) = \pi(\phi_1)\pi(\omega_1)\pi(\alpha_{11})\pi(\alpha_{12})$ where $\phi_1 \sim N(0, 0.1)$, $\omega_1 \sim U(0, \infty)$, $\alpha_{11} \sim U(0, 1)$ and $\alpha_{12} \sim U(0, 1)$.

For the *Expert 2*, $\eta_2 = (\phi_2, \omega_2, \alpha_{21}, \alpha_{22})$ also has components that are a priori independent where, $\phi_2 \sim N(0, 0.1)$, $\omega_2 \sim U(0, \infty)$, $\alpha_{21} \sim U(0, 1)$ and $\alpha_{22} \sim U(0, 1)$.

In an analogous way, for *Expert 3* the vector $\eta_3 = (\phi_3, \omega_3, \alpha_{31}, \alpha_{32}, \alpha_{33})$ has independent components with marginal prior distributions given by $\phi_3 \sim N(0, 0.1)$, $\omega_3 \sim N(0, 10)$, $\alpha_{31} \sim U(-1, 1)$, $\alpha_{32} \sim N(0, 10)$ and $\alpha_{33} \sim N(0, 10)$.

Finally, each of the entries of the vector of parameters appearing in the mixture weights, $\gamma = (u_1, u_2, v_1, v_2)$, is assumed to have a $U(-l, l)$ prior distribution with a large value for l. This prior specification was chosen to reflect vague (flat) prior information and to facilitate the calculations of the different steps inside our MCMC scheme. The $N(0, 0.1)$ prior on the AR(1) coefficients was proposed with the idea of containing most of its mass in the region defined by the stationarity condition. Instead of a $N(0,0.1)$ prior, we also used a $U(—1, 1)$ prior on the coefficients and the results obtained were essentially the same as with the Normal prior. In fact, the noninformative priors were suggested by Vrontos et al. (2000) in the context of parameter estimation, model selection and volatility prediction. Also, these authors show that under these priors and for pure GARCH/EGARCH models, the difference between classical and Bayesian point estimation is minimal. The restrictions on these priors in the different parameter spaces is to satisfy the stationarity conditions of the expert models. For our ME model, these type of noninformative priors were key to produce good MCMC convergence results which could not be obtain with other classes of (informative) priors. Since the parameters of ARCH/GARCH/EGARCH models are not very meaningful in practice, the most typical prior specification for these parameters is to adopt noninformative prior distributions. To our knowledge, there is no study on the effects of using informative priors in this context. Furthermore, Villagran (2003) discusses another aspect of our prior specification in terms of simulated data. If the true data follow an AR(1)-GARCH(1,1) structure, the noninformative prior allows to estimate the parameters of the true model with a maximum absolute error of 0.003. The maximum posterior standard deviation for all the model parameters is 0.0202 and the posterior mean for $g_t^{(t)}$ is practically equal to one for the true model (in this case GARCH) and for all time t.

Our MCMC algorithm can be summarized as follows:

- Assign initial values for θ and with these values calculate the volatilities for each expert, $\sigma_{1,t}^{2(0)}$, $\sigma_{2,t}^{2(0)}$ and $\sigma_{3,t}^{2(0)}$ for all t.

- Evaluate the probabilities $g_r^{(t)}$ at $\gamma^{(0)}$ and compute the conditional probabilities $h_r^{(t)}$ for all t.
- Generate $z_r^{(t)}$ conditional on θ, $\mathscr{F}_n$, χ from a Multinomial distribution with total count equal to 1 and cell probabilities $h_r^{(t)}$. Across this step, we are generating vectors $(z_1^{(t)}, z_2^{(t)}, z_3^{(t)})$ for all values of t.
- For the augmented posterior distribution, we generate each of the expert parameters and each of the mixture weights or gating function parameters via Metropolis–Hastings (M–H) steps. A general description of the M–H algorithm with several illustrative examples appears in Tanner (1996).
- At each M–H step, we propose a new value $\theta^{(j)}$ for the parameters from a candidate distribution and than accept or reject this new value with probability

$$\alpha(\theta^{(j-1)}, \theta^{(j)}) = min\left[\frac{\pi(\theta^{(j)}|\cdot)q(\theta^{(j-1)})}{\pi(\theta^{(j-1)}|\cdot)q(\theta^{(j)})}, 1\right].$$

- After generating all the model parameters at iteration j, we update the volatilities $\sigma_{1,t}^{2(j)}$, $\sigma_{2,t}^{2(j)}$ and $\sigma_{3,t}^{2(j)}$, the probabilities $g_r^{(t)}$, $h_r^{(t)}$ and the indicator variables $\boldsymbol{Z}^{(j)}$.
- The algorithm is iterated until Markov Chain convergence is reached. An initial section of the iterations is considered a burn-in period and the remaining iterations are kept as posterior samples of the parameters.

Given that mixture models are highly multimodal, to improve on the convergence of our MCMC method, it is convenient to propose several starting points for θ and run the algorithm for a few iterations. The value that produces the maximum posterior density is used as the initial point $\theta^{(0)}$ to produce longer runs of the Markov chain. In the applications that are presented in the next section, we used 20 overdispersed starting values for θ and calibrated our proposal distribution so that the acceptance rates of all the M-H steps were around 45%. We maintained this acceptance rates relatively low to allow full exploration of the parameter space and to avoid getting stuck around a local mode.

For a specific MCMC iteration, the volatility or conditional variance of our ME model is computed as

$$V_t^{(j)} = \sum_{r=1}^{3} g_{r,t}^{(j)} \sigma_{r,t}^{2(j)} + \sum_{r=1}^{3} g_{r,t}^{(j)} (\mu_{r,t}^{(j)} - \bar{\mu}_t^{(j)})^2$$

$$\bar{\mu}_t^{(j)} = \sum_{r=1}^{3} g_{r,t}^{(j)} \mu_{r,t}^{(j)} = \sum_{r=1}^{3} g_{r,t}^{(j)} \phi_r^{(j)} y_{t-1}$$

where the index t represents time, the index j the iteration and $\mu_{r,t}$ the mean of expert r at time t. These expressions follow from well-known results to compute

the variance of a mixture distribution with three components. Given a value of model parameters at iteration j, $V_t^{(j)}$ can be directly evaluated from these equations. The expression for the volatility of our ME model is formed by two terms. The first term represents the dependency of the conditional variance with respect to past volatilities, and the second term represents changes in volatility of the mixture model due to the differences in conditional mean between experts.

In the next section, we show that using time as a covariate allows one to detect structural changes in volatility so ME is able to determine if the process generating the data corresponds to a unique expert.

4. APPLICATIONS

4.1. Exchange Rates US Dollar/German Mark

Figure 1(a) shows 500 daily observations between the American dollar and the German mark starting on October of 1986 and Fig. 1(b) shows the corresponding returns of these exchange rates.[1]

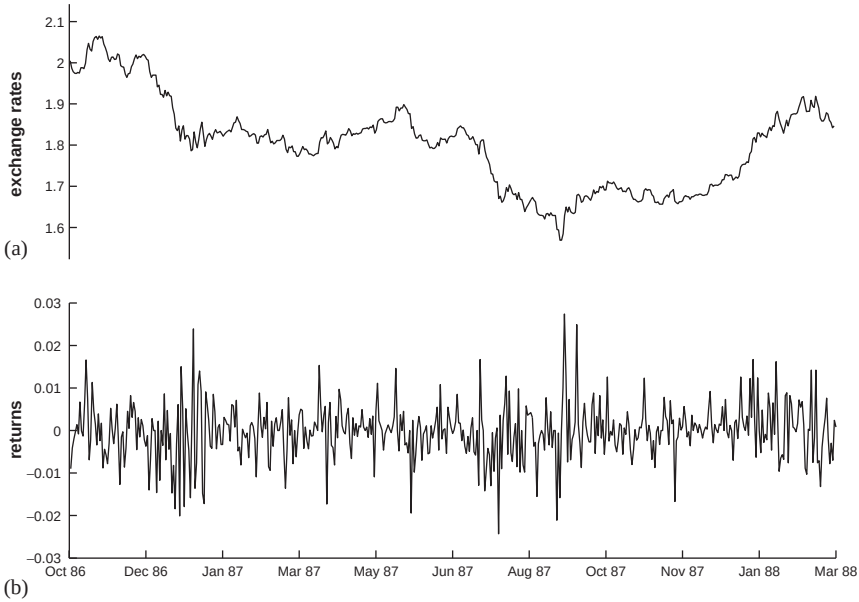


Fig. 1. (a) Exchange Rates between U.S. Dollar and German Mark Starting from October 1986. (b) Returns of the Exchange Rates.

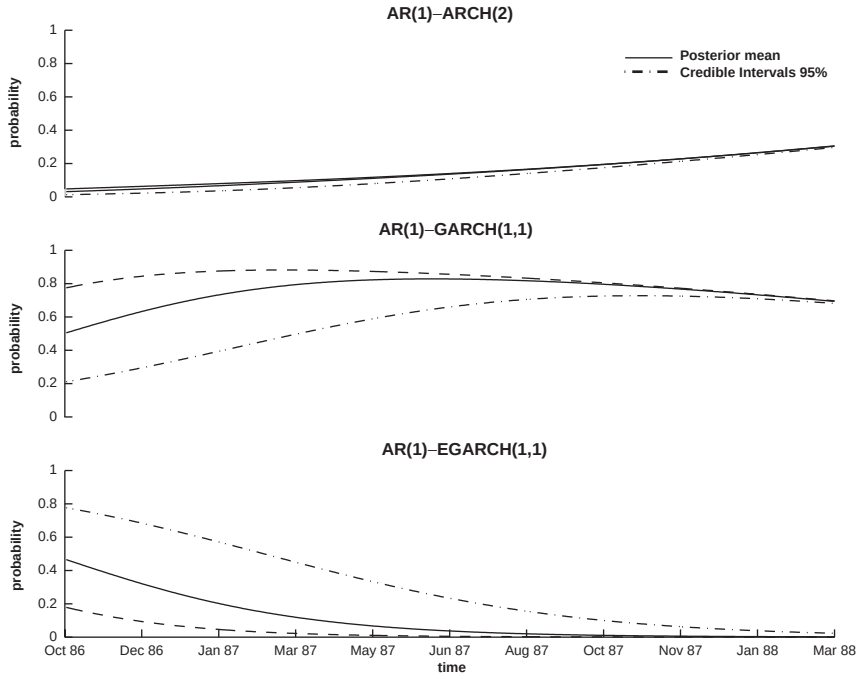


Fig. 2. Exchange Rates Example. Posterior Means of $g_r^{(t)}; r = 1, 2, 3$ (Solid Lines) and 95% Credible Intervals (Dashed Lines).

Using the returns as our response time series, we implemented the ME model with time being our covariate. Figure 2 shows posterior means and 95% credible intervals for the unconditional probabilities of each expert, i.e., $g_r^{(t)}; r = 1, 2, 3$. Both $h_r^{(t)}$ and $g_r^{(t)}$ are functions of the unknown parameter vector θ, so it makes absolute sense to assess measures of uncertainty to these probabilities.

We can appreciate that the expert that dominates in terms of probability is the AR(1)-GARCH(1,1) model. Furthermore, Fig. 2 also shows the relative uncertainty of the different experts across time. For the period covering October 1986 to January 1987, there is a significant weight associated to the AR(1)-EGARCH(1,1) model and the credible band for the weight of this model can go as high as 0.8 and as low as 0.2. In Fig. 3, we report posterior means of the conditional probabilities of each model, $h_r^{(t)}; r = 1, 2, 3$.

This figure shows a similar pattern compared to the description of probabilities given by Fig. 2. Since these are conditional probabilities, individual

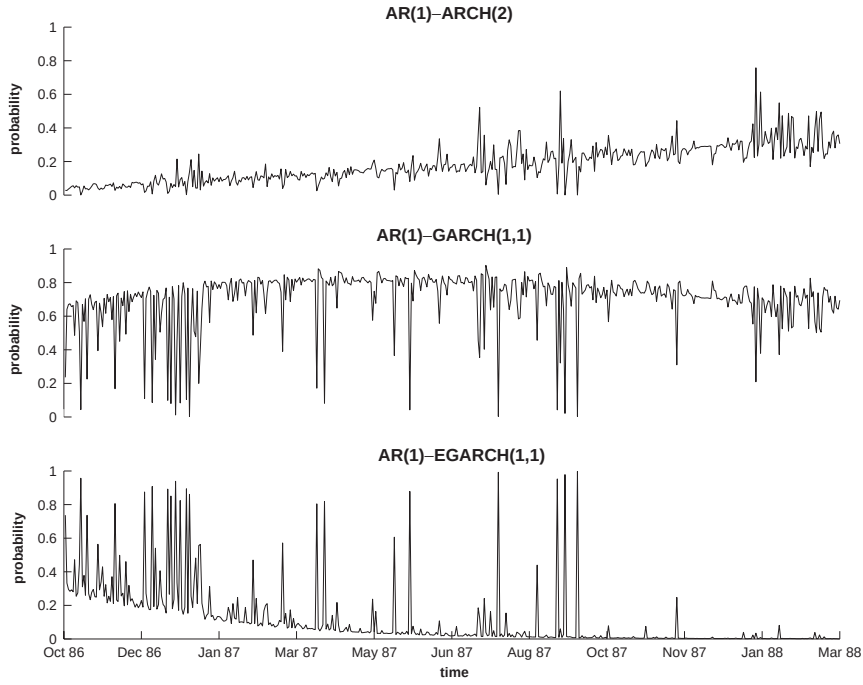


Fig. 3. Exchange Rates Example. Posterior Means of $h_r^{(t)}; r = 1, 2, 3$.

observations may produce high fluctuations in probability. However, this figure confirms that the models that dominate in the ME, at least in the initial time periods, are the AR(1)-GARCH(1,1) and the AR(1)-EGARCH(1,1). Toward the end of the considered time periods, the dominant model is the AR(1)-GARCH(1,1) but the AR(1)-ARCH(2) model has a significant weight of 0.4.

In Fig. 4, we present posterior mean estimators of the volatility for the ME model and for the individual expert models.

The ME model has a higher volatility estimate at the beginning of the time series in contrast to the individual models. This is due to the influence of the EGARCH models in the first part of the series as shown by Figs. 2 and 3. A referee and the editors suggested that we compared the square of the residuals of a pure AR(1) model fitted to the return series, with the volatilities presented in Fig. 4. These residuals seem to be better charaterized by the AR(1)-ARCH(2) volatilities towards the end of the time period covered by the data. However at the beginning of the period, the residuals are

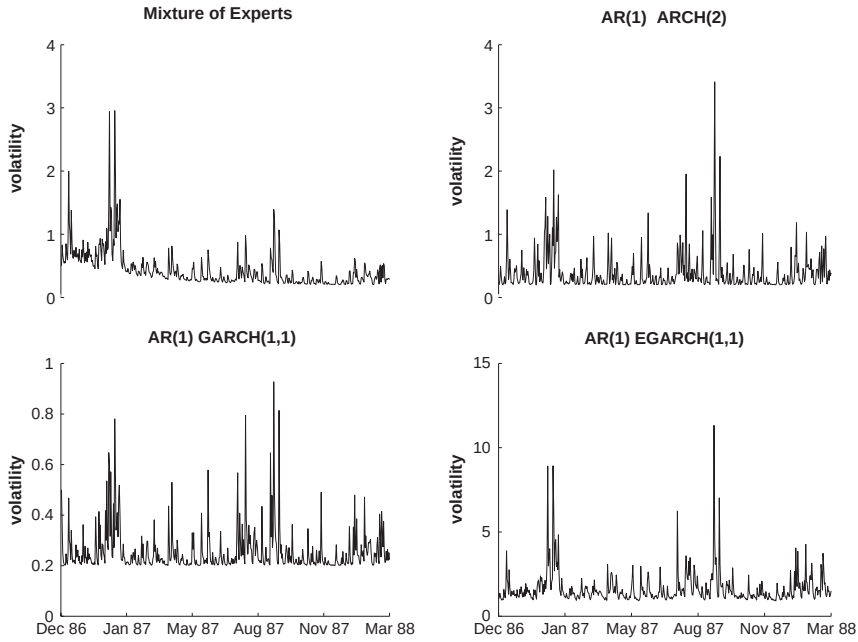


Fig. 4. Exchange Rates Example. Posterior Mean Estimate of Volatility for Mixture-of-Experts Model and Posterior Estimates of Volatility for Individual Models.

more closely followed by the AR(1)-EGARCH(1,1) volatilities. Additionally, we think that the ME is at least as good as any individual expert model since it is pooling information from different models. A great advantage of the ME model is that it shows, as a function of time *t,* how different expert models are competing with each other conditional on the information at $t - 1$ and how the volatilities change according to time. Notice that from January 1987, the volatilities of the AR(1)-ARCH(2) are consistent with the volatilities of the ME model.

Figure 5 considers a ''future'' period starting from March 1988 and that covers 100 daily exchange-rate values.

Figure 5(a) shows the time series of returns for this future period and Fig. 5(b) shows the one-step-ahead predictive posterior means and the 95% one-step-ahead forecast intervals for volatility based on the ME model that only uses previous data from October 1986 to March 1988 and with a forecasting horizon of 100 time steps. This figure illustrates one of the main

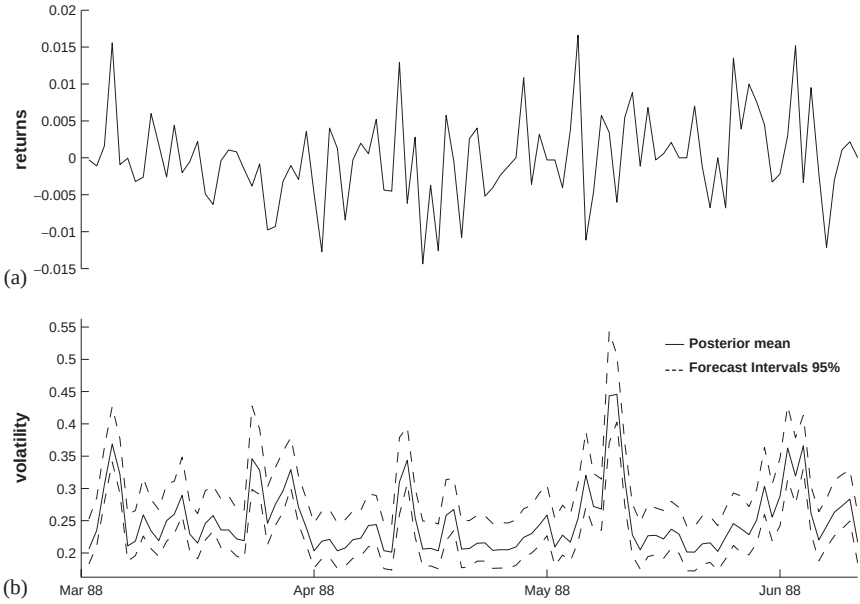


Fig. 5. Exchange Rates Example. (a) Time Series of Returns Starting from March 1988. (b) Predictive Posterior Means and 95% Forecast Intervals for Volatility.

features of our model. The MCMC approach allows us to compute samples of future or past volatilities values that can be summarized in terms of predictive means and credible intervals. In our ME model, the volatility is a very complicated function of the parameters and producing non-Monte Carlo estimates, especially predictive intervals, is practically impossible.

4.2. Analysis of the Mexican Stock Market

In this application, we studied the behavior of the Mexican stock market (IPC) index using as covariates the Dow Jones Industrial (DJI) index from January 2000 to September 2004 and also using time. Figure 6 shows both the IPC index and the DJI index time series with their corresponding returns.

It is obvious that the IPC index and the DJI index have the same overall pattern over time. In fact, some Mexican financial analysts accept that the IPC index responds to every 'strong' movement of the DJI index. Our Bayesian ME approach adds some support to this theory.

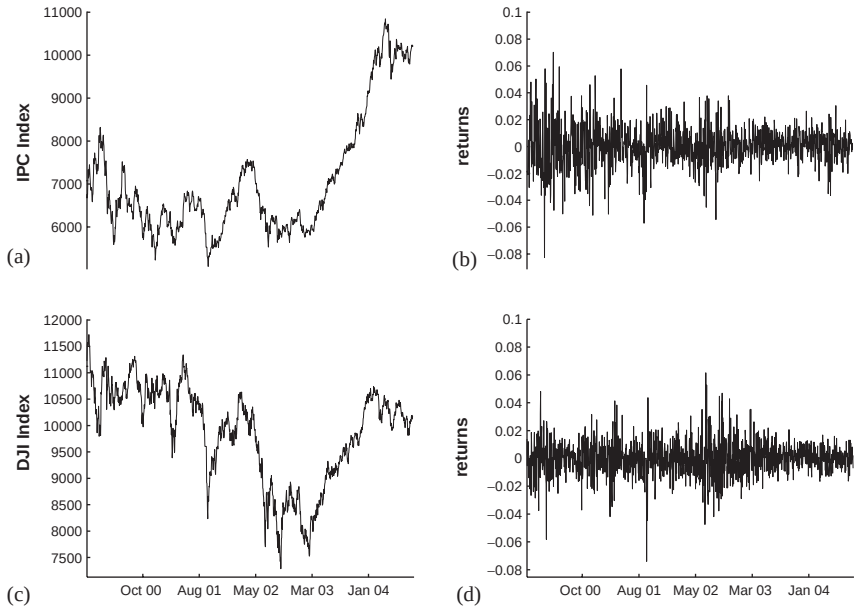


Fig. 6. (a) The Mexican Stock Market (IPC) Index from January 2000 to September 2004. (b) Time Series of Returns for the IPC Index. (c) The Dow Jones Index (DJI) from January 2000 to September 2004. (d) Time Series of Returns for the DJI.

In Fig. 7, we show the posterior distributions of the parameters for the mixture weights or gating functions when W_t was set equal to the DJI index. The posterior distribution for the 'slope' parameters u_1 and u_2 have most of their posterior mass away from 0, which means that the effect of the covariate in our ME analysis is 'highly significant'.

In Fig. 8 we show the mixture weights of the ME using different covariates.

The left column presents the posterior mean estimates of $g_r(t); r = 1, 2, 3$ using the DJI index as covariate and the right column shows the estimates as a function of time. The right column shows a shift on the regime since the AR(1)-EGARCH(1,1) expert rules the evolution of the volatility of the IPC index from January 2000 to March 2002. After this date, the AR(1)-ARCH(2) model is the one with higher probability. As a function of the DJI index, the mixture weights behavior is quite different. Now the AR(1)-EGARCH(1,1) expert has higher probability in comparison to the other experts. This is a result due to the common high volatility shared by the IPC index and the DJI index.

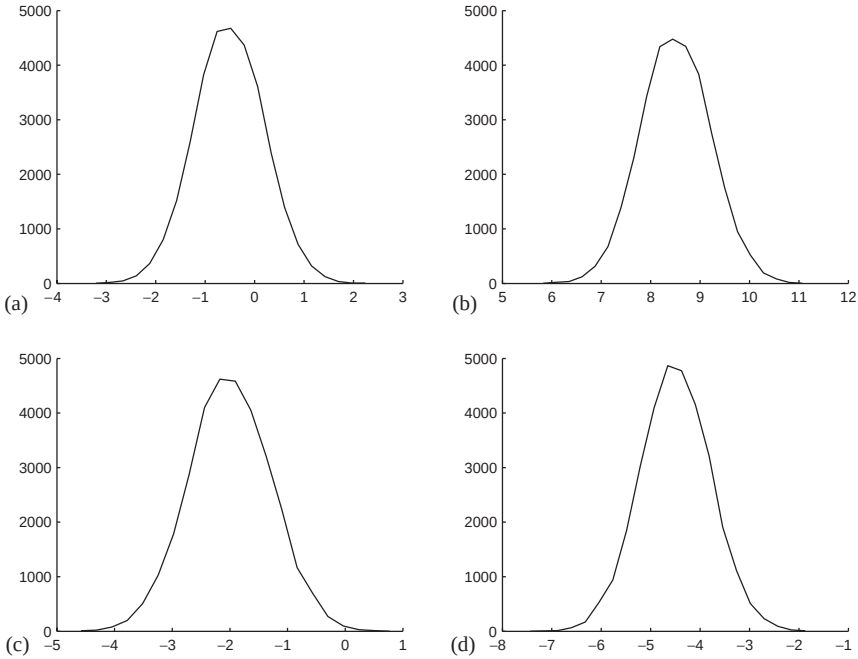


Fig. 7. (a) Posterior Distribution for Gating Parameter v_1. (b) Posterior Distribution for Gating Parameter u_1. (c) Posterior Distribution for Gating Parameter v_2. (d) Posterior Distribution for Gating Parameter u_2.

5. CONCLUSIONS AND EXTENSIONS

In this paper, we present a mixture modeling approach based on the Bayesian paradigm and with the goal of estimating SV. We illustrate the differences of our mixture methodology versus a sole model approach in the context of ARCH/GARCH/EGARCH models for two different financial series. The two main aspects of our ME model are: (1) the comparison of different volatility models as a function of covariates and (2) the estimation of predictive volatilities with their corresponding measure of uncertainty given by a credible interval. On the other hand, we had only considered ME and not HME. The difficulty with HME is that it requires the estimation of the number of overlays O, which poses challenging computational problems in the form of reversible jump MCMC methods. Additionally, we had not considered any other experts beyond ARCH, GARCH and EGARCH

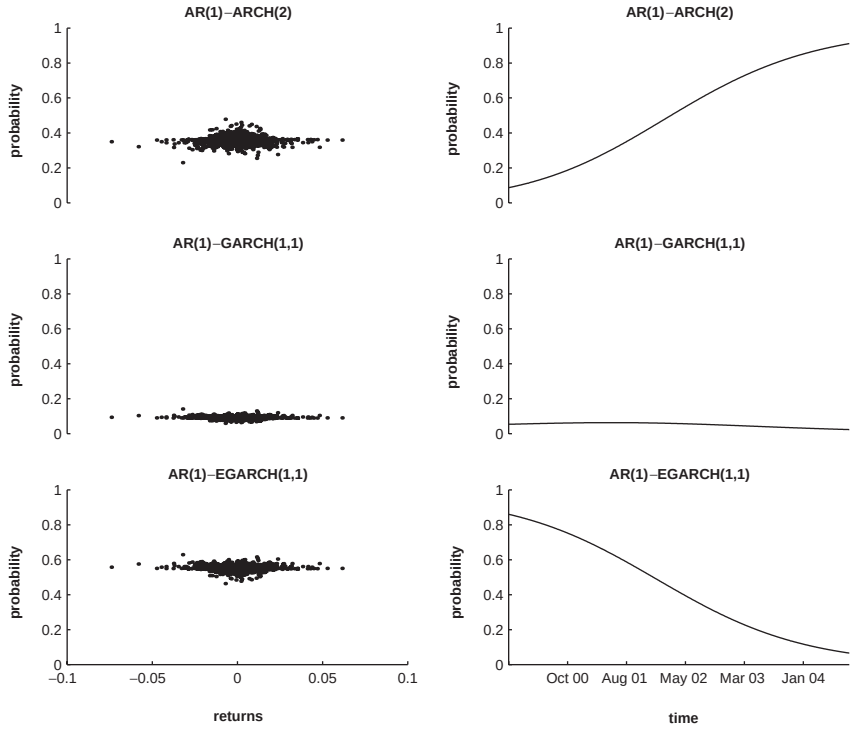


Fig. 8. Left Column. Probabilities of Each Expert as a Function of the Returns of the DJI. Right Column. Probabilities of Each Expert as a Function of Time.

models. An extension to our approach considers other competitors or experts like the SV models of Jacquier, Polson, and Rossi (1994). This leads into MCMC algorithms combining mixture modeling approaches with *Forward filtering backward simulation.* These extensions are part of future research.

NOTE

1. The code to fit the models used for this section is available under request from avhstat@unm.edu. Also, this code can be downloaded from http://www.stat.unm.edu/~avhstat.

ACKNOWLEDGMENTS

We wish to express our thanks to Professors Thomas Fomby and Carter Hill, editors of this volume, for all their considerations about this paper. During the preparation of this paper, A. Villagran was partially supported by CONACyT-Mexico grant 159764 and by The University of New Mexico.

REFERENCES

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

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*. New York: Oxford University Press.

Huerta, G., Jiang, W., & Tanner, M. A. (2001). Mixtures of time series models. *Journal of Computational and Graphical Statistics*, *10*, 82–89.

Huerta, G., Jiang, W., & Tanner, M. A. (2003). Time series modeling via hierarchical mixtures. *Statistica Sinica*, *13*, 1097–1118.

Jacquier, E., Polson, N., & Rossi, P. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, *12*, 371–389.

Jordan, M., & Jacobs, R. (1994). Hierarchical mixture of experts and the EM algorithm. *Neural Computation*, *6*, 181–214.

Melino, A., & Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometric*, *45*, 239–265.

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

Peng, F., Jacobs, R. A., & Tanner, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, *91*, 953–960.

Tanner, M. A. (1996). *Tools for statistical inference* (3rd ed.). New York: Springer-Verlag.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–550.

Tsay, R. S. (2002). *Analysis of financial time series*. New York: Wiley.

Villagran, A. (2003). *Modelos Mezcla para Volatilidad*. Unpublished MSc. thesis, Universidad de Guanajuato, Mexico.

Vrontos, D., Dellaportas, P., & Politis, D. N. (2000). Full Bayesian inference for GARCH and EGARCH models. *Journal of Business & Economic Statistics*, *18*, 187–197.

Wong, Ch. S., & Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association*, *96*, 982–995.

A MODERN TIME SERIES ASSESSMENT OF "A STATISTICAL MODEL FOR SUNSPOT ACTIVITY" BY C. W. J. GRANGER (1957)

Gawon Yoon

ABSTRACT

In a brilliant career spanning almost five decades, Sir Clive Granger has made numerous contributions to time series econometrics. This paper reappraises his very first paper, published in 1957 on sunspot numbers.

1. INTRODUCTION

In 1957, a young lad from the Mathematics Department of Nottingham University published a paper entitled "*A Statistical Model for Sunspot Activity*" in the prestigious *Astrophysical Journal*, published nowadays by the University of Chicago Press for the American Astronomical Society. The paper showed that sunspot numbers could be well represented by a simple statistical model "formed by the repetition of a fixed basic cycle varying only in magnitude and phase." The paper was received by the editorial office of the *Journal* on the author's 22nd birthday and is still cited, see for instance

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 297–314
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20031-2

Reichel, Thejll, and Lassen (2001) and Carozza and Rampone (2001), even though the author himself stopped working on the sunspot numbers quite sometime ago. The author had just received a bachelor's degree in mathematics and was an assistant lecturer at the University while he was still a Ph.D. student there. He did not know that there was a submission fee to the *Journal*, which was kindly waived by its editorial office.[1] The author of the paper is Clive W. J. Granger, a Nobel Memorial Prize winner in Economic Science in 2003. His contributions to time series econometrics are numerous and far-reaching, and they are not easy to summarize in brief; see some of his work collected in Granger (2001) as well as about a dozen of his books. Hendry (2004) provides a useful summary of Granger's major contributions to time series analysis and econometrics.[2] From an interview conducted by Peter Phillips for *Econometric Theory* in 1997, one may gather that the young Granger was initially interested in meteorology; see Granger (1997). Interestingly, his second published paper in 1959 was concerned with predicting the number of floods of a tidal stretch; see Granger (1959). However, most of his later work is related to economic data, especially time series data.

It is not yet widely acknowledged that Granger's early work on sunspots had also led to an important theoretical development in economics. Karl Shell, who formulated the concept of sunspot equilibrium with David Cass, notes during his visit to Princeton that "Clive gave me some of his technical reports, which contained reviews of time-series literature on the possible economic effects of real-world sunspots *à la* Jevons. Perhaps this was the seed that germinated into the limiting case of purely extrinsic uncertainty, i.e., stylized sunspots *à la* Cass-Shell." (See Shell, 2001).[3]

Granger (1957) proposed a simple two-parameter mathematical model of sunspot numbers, with an amplitude factor and an occurrence of minima, respectively. He reported that his model accounted for about 85% of the total variation in the sunspot data. He also made several interesting observations on sunspot numbers. It is the aim of this paper to determine if the statistical model he proposed stands the test of the passage of time with the extended sunspot numbers that span beyond the sample period he considered.[4] In the next Section, a brief discussion on sunspot data is provided.

2. DATA ON SUNSPOT NUMBERS

Sunspots affect the weather as well as satellite-based communications. They are also suspected of being responsible for global warming; see Reichel et al.

(2001). Sunspot numbers have been recorded for more than 300 years and are known to be very difficult to analyze and forecast. For instance, Tong (1990) notes that they "reveal an intriguing cyclical phenomenon of an approximate 11 year period which has been challenging our intellect…."

Efforts in solar physics have been made over the years to understand the causes of sunspot cycles. For instance, Hathaway, Nandy, Wilson, and Reichman (2003, 2004) recently show that a deep meridional flow, a current of gas supposed to travel from the poles to the equator of the Sun at a depth of about 100,000 km,[5] sets the sunspot cycle. Given that various theoretical models for sunspot cycles are still in their early stages of development, time series models appear to be a natural alternative in modeling sunspot numbers. Indeed, various time series models are already applied to them, such as linear autoregressive, threshold, and bilinear models, among many others. In fact, sunspot numbers were the first data to be analyzed by autoregressive models: in 1927, Yule fitted an AR(2) model to sunspot data. See also Section 7.3 of Tong (1990) for some estimation results from various time series models. Granger also came back to the sunspot numbers later in Granger and Andersen (1978), using his bilinear model. Forecasting sunspot cycles is not easy; for instance, see Morris (1977) and Hathaway, Wilson, and Reichman (1999).

In the remainder of this Section, the sunspot data used in this paper will be presented and compared with those employed in Granger (1957).[6] G(57) used both annual and monthly sunspot numbers from 1755 to 1953, available from *Terrestrial Magnetism and Atmospheric Electricity* and *the Journal of Geophysical Research*. The annual sunspot numbers he used are listed in his Table 2. G(57) also contained a brief discussion on how the sunspot numbers are constructed. In this paper, the extended annual sunspot numbers are employed, which are available from the Solar Influences Data Analysis Center (SIDC), http://sidc.oma.be/DATA/yearssn.dat, for the sample period of 1700–2002.[7] Figure 1 shows the annual sunspot numbers. The variation in amplitude and period can easily be seen. It turns out that for the sample period of 1755–1953, the extended annual sunspot numbers are not the same as those used in G(57). Figure 2 shows the differences between the two sets of sunspot numbers. The biggest difference occurs in 1899; 12.1 in this paper and 21.1 in G(57). The discrepancy is most likely due to a type-in error. There is also a minor error in G(57) in calculating the mean for the cycle that started in 1856; in his Table 2, the mean should be 49.6, not 51.7. As this wrong mean value was employed in the ensuing calculations, the results used in G(57) for his Figs. 1 and 2 are not correct;

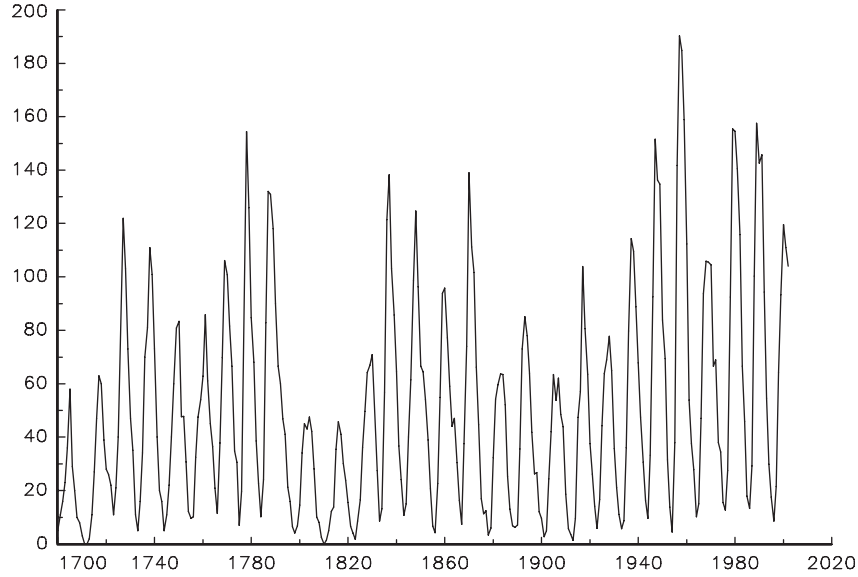


Fig. 1. Annual Sunspot Numbers, 1700–2002.

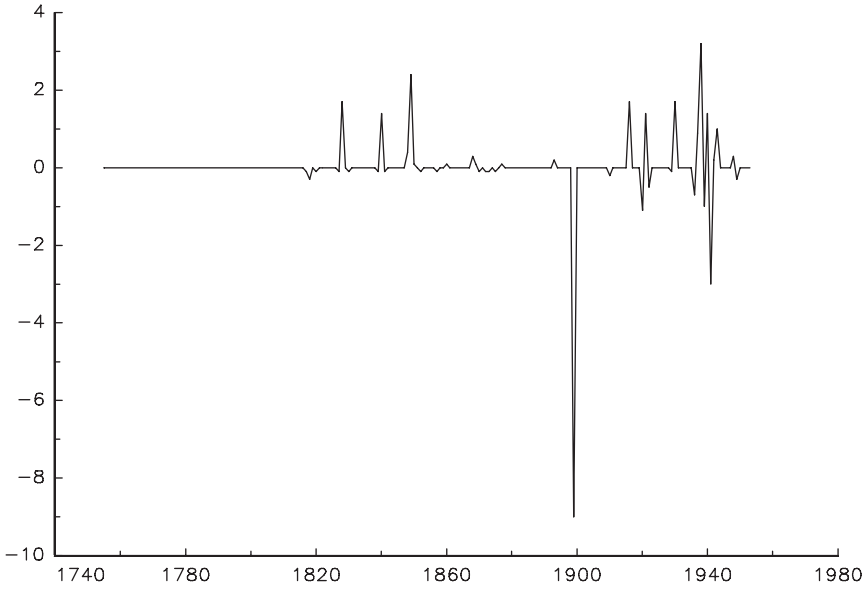


Fig. 2. Differences Between the Two Sets of Data, 1755–1953.

Table 1. Sunspot Cycles with Annual Observations.

Cycle No.	Year of Minimum	Period (Years)	Mean	From Granger (1957)
1*	1700	12	18.3	
2	1712	11	29.5	
3	1723	10	54.1	
4	1733	11	51.5	
5	1744	11	40.1	
6	1755	11	42.4	42.4
7	1766	9	59.9	60.0
8	1775	9	68.2	68.2
9	1784	14	60.45	60.45
10	1798	12	23.8	23.8
11	1810	13	18.1	18.1
12	1823	10	39.0	38.9
13	1833	10	65.3	65.2
14	1843	13	53.7	53.5
15	1856	11	49.6	51.7 [49.6]***
16	1867	11	56.6	56.6
17	1878	11	34.6	34.6
18	1889	12	38.8	39.5
19	1901	12	31.1	31.1
20	1913	10	44.2	44.0
21	1923	10	41.0	40.9
22	1933	11	55.0	54.8
23	1944	10	75.7	75.7
24	1954	10	95.0	
25	1964	12	58.8	
26	1976	10	82.9	
27	1986	10	78.5	
28**	1996	NA	NA	

NA: Not applicable.
*The first cycle is assumed to start in 1700, which is the first year the sunspot numbers became available. Granger (1957) instead called the cycle that started in 1755 Cycle 1.
**The last cycle, Cycle 28, is not yet completed.
***The number in the square brackets, 49.6, is the correct mean, not 51.7, which was originally reported in Granger (1957).

but the effect is only minor. Compare also the last two columns of Table 1 below: Other than the error for the mean of the cycle that started in 1856, the differences in means are due to the fact that the two sunspot number data sets are not the same. The main empirical results of this paper are presented in the next Section.

3. MAIN EMPIRICAL RESULTS

G(57) made several interesting observations on sunspot activity and provided supporting evidence. In this Section, empirical regularities discussed in G(57) are reappraised with the extended sunspot numbers. Of course, properties of sunspot cycles other than those discussed in G(57) are also available in the literature; see Hathaway, Wilson, and Reichman (1994). The organization of this Section follows closely that of G(57) to make comparison easier.

3.1. Varying Periods

In this paper, the sunspot cycle that started in 1700, the first year in which the annual sunspot numbers become available, will be called Cycle 1. G(57) instead called the cycle that started in 1755 Cycle 1. In Table 1, the date for the start of sunspot cycles, their periods, and the two means for the two different sunspot numbers are listed for the 27 (and 18) completed cycles. Table 1 shows that periods are changing, as already noted in G(57). The mean of periods is 11.0 years for the extended sunspot numbers, with a standard deviation of 1.2 years. For the sunspot numbers used in G(57), they are 11.1 and 1.4 years, respectively.

3.2. The Basic Cycle

G(57) assumed that sunspot cycles are "made up of a basic cycle, each multiplied by an amplitude factor plus a random element with zero mean." See Eq. (4) for his model. The basic cycle was determined as follows: "The mean of the sunspot numbers for the years making up the cycle was found, and all the yearly intensities were divided by this mean. The new numbers were then averaged for the minimal year, the year after the minimum, and so forth over all the cycle" (p. 153).[8] Figure 3 shows the newly transformed annual sunspot numbers. They display similar amplitudes, except for Cycle 1. Figure 4 displays the basic cycles from the two different sunspot number data sets, with the time from the minimum on the horizontal axis.[9] The two basic cycles are very close to each other, except at 5 years after the minimum. For the extended sunspot data, the basic cycle reaches a maximum in 5 years after the minimum, not in 4 years as in the sunspot data used in G(57). However, it is still true that it takes less time to get to the maximum from the minimum than from the maximum to the next minimum,

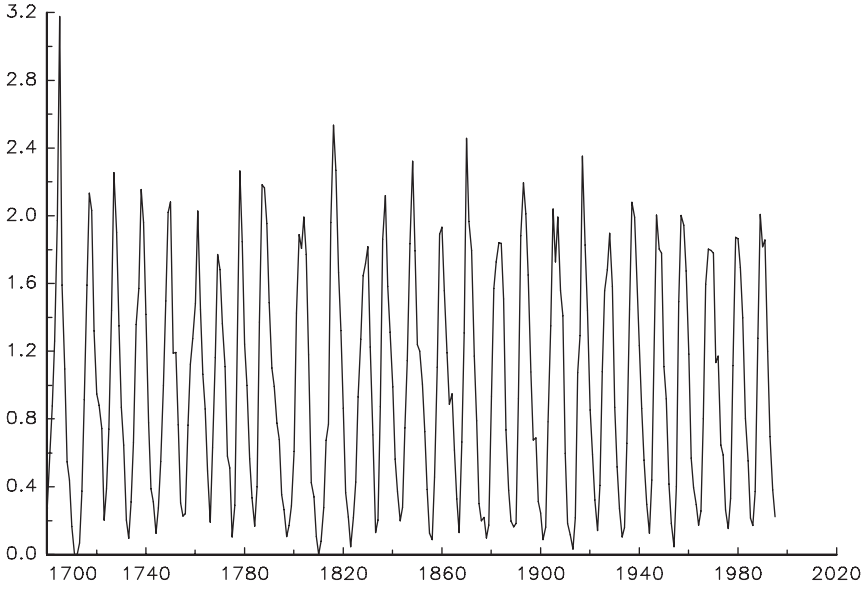


Fig. 3. Transformed Annual Sunspot Numbers, 1700–1995.

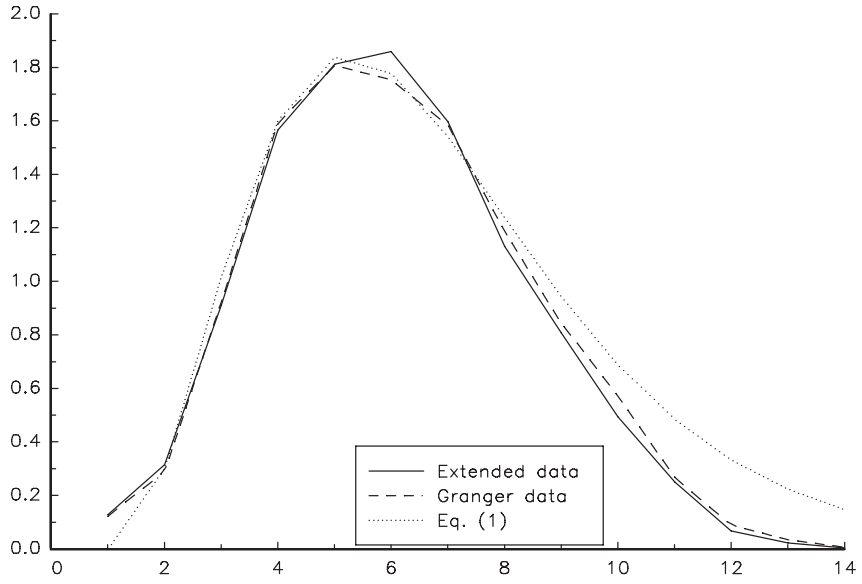


Fig. 4. The Basic Cycle, Corresponding to Fig. 1 in Granger (1957).

as already observed in G(57). G(57) also proposed the following functional form to approximate the basic curve:

$$g = t^{2.67}\mathrm{e}^{1.73-0.63t} \tag{1}$$

where t denotes the time from the minimum and g the intensity. (1) belongs to the model considered by Stewart and Panofsky (1938) and Stewart and Eggleston (1940).[10] Figure 4 shows that (1) fits the basic curves rather well, especially over the middle range.

3.3. Long-Term Fluctuations

G(57) also examined the cycle means over time. The real line in Fig. 5 shows the means of each cycle for the sunspot numbers employed in G(57).[11] It appears that the means tend to increase, especially for the later cycles. In Fig. 5, the following relation representing a long-term fluctuation is also superimposed:

$$46 + 19 \times \sin\left(\frac{360}{87}t + 7.25\right) \tag{2}$$

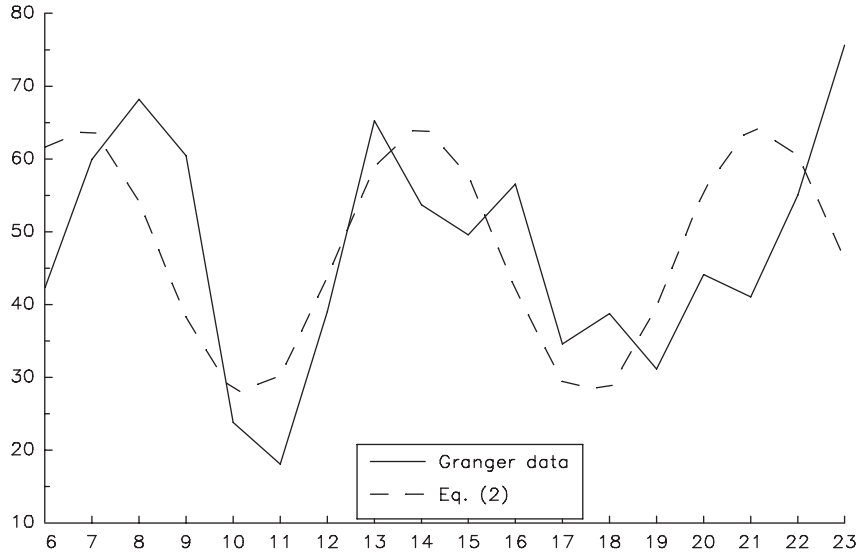


Fig. 5. Means of Cycle 6–23, Corresponding to Fig. 2 in Granger (1957).

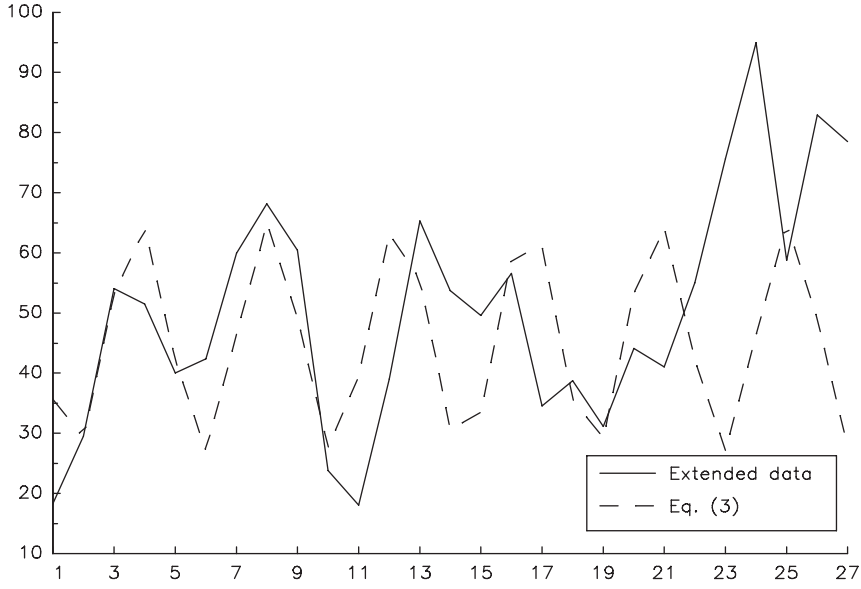


Fig. 6. Means of Cycle 1–27.

as proposed in G(57). The relation (2) seems to track the changes in means of sunspot cycles rather well, except for the later cycles. The correlation between the two curves plotted in Fig. 5 is 0.52, which is slightly lower than that reported in G(57). In Fig. 6, the cycle means are plotted with a real line for the extended sunspot numbers. Clearly, the means of Cycles 24–27 become higher. Therefore, it easily follows that the model of long-term fluctuation in the form of (2) would perform less well for the extended sunspot number data. For instance, the following model of long-term sunspot fluctuation is also plotted in Fig. 6, with a broken line:

$$46 + 19 \times \sin\left(\frac{360}{87}t + 10.0\right) \tag{3}$$

which produces a correlation of only 0.18 with the basic curve.[12]

3.4. Goodness of Fit of the Model

The model G(57) proposed is of the form for sunspot activities

$$f(x)\{g(x) + \varepsilon_x\} \tag{4}$$

where $f(x)$ is the amplitude factor, shown in Figs. 5 and 6, $g(x)$ the basic curve in Fig. 4, and ε_x a zero-mean random variable.[13] The two-parameter model of sunspot activity accounted for about 86% of the total variation for the data G(57) used. For the extended sunspot data with 27 completed cycles, the fit of his model is still high, explaining about 82% of the total variation. Hence, the model Granger proposed in 1957 still appears to be effective in accounting for the fluctuations in the extended sunspot numbers as well as in the data series he employed.

3.5. Additional Observations

G(57) noted that mean solar activity is negatively correlated with sunspot cycle length: "the cycle is often shorter, the higher the mean or peak amplitude." He employed a 2×2 contingency table to test the claim. In this paper, a simple linear regression model will be tested instead. Figure 7 shows that there is indeed an inverse relation between the mean and the period (in years) of a cycle.[14] For instance, the following estimation result is found with OLS for the 27 cycles from the extended sunspot data:

$$\underset{}{\text{mean}} = \underset{(31)}{131} - \underset{(2.81)}{7.33}\,\text{period}$$

with $R^2 = 0.21$. Standard errors are reported in the parentheses. The estimated coefficient associated with *period* is significant at conventional significance levels. The fitted line is also shown in Fig. 7. For the sunspot numbers used in G(57), the results are

$$\underset{}{\text{mean}} = \underset{(28)}{98} - \underset{(2.44)}{4.53}\,\text{period}$$

with $R^2 = 0.17$ and sample size $= 18$. Additionally, using the functional specification discussed in G(57), the following estimation results are obtained. For the extended sunspot data,

$$\underset{}{\text{mean}} = \underset{(31.4)}{-38.5} + \underset{(339)}{966}\,\text{period}^{-1}, \quad R^2 = 0.25$$

For the sunspot data used in G(57),

$$\underset{}{\text{mean}} = \underset{(28.5)}{-10.2} + \underset{(308)}{630}\,\text{period}^{-1}, \quad R^2 = 0.21$$

It should also be added, however, that there is only a very weak relation between the mean and the period for the monthly data, to be discussed below.

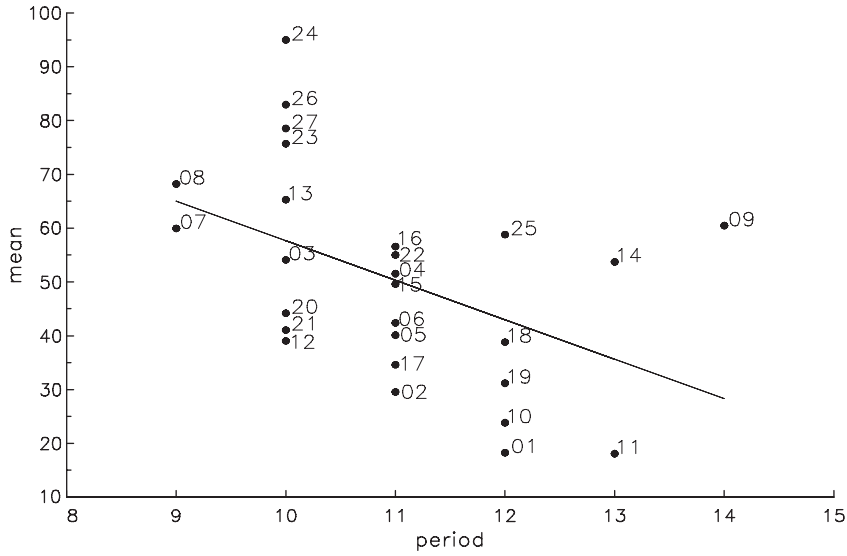


Fig. 7. Scatter Diagram Between Cycle Period and Mean of 27 Cycles.

G(57) contained further discussions on the sunspot numbers. The following results are obtained from monthly sunspot numbers, which are available at http://sidc.oma.be/index.php3. More specifically, *smoothed* monthly sunspot numbers for the sample period of 1749:07–2003:05 are used in this paper.[15] Figure 8 plots the monthly sunspot numbers. Clearly, they are less smooth than the annual ones. Table 2 shows the starting date for each sunspot cycle with its period in months and mean sunspot number. The mean and standard deviation of periods are 131.5 and 14.3 months, respectively, for the 22 completed cycles. Unfortunately, the monthly sunspot numbers used in G(57) are not available to this author.

G(57) noted that "the peak came later for cycles of small mean as compared to cycles with larger means." This relation is known in the literature as the Waldmeier effect. G(57) defined *peak* by the time in months between the minimum and the maximum of a cycle divided by its length. Figure 9 shows a scatter plot of the *peak* and *mean* of sunspot cycles; note that the first cycle observed for the monthly sunspot numbers is Cycle 6, as listed in Table 2. Figure 9 seems to indicate a negative relation between the *peak* and *mean* of sunspot cycles. For instance, for the 22 completed sunspot cycles from the monthly observations, the following estimation results are found

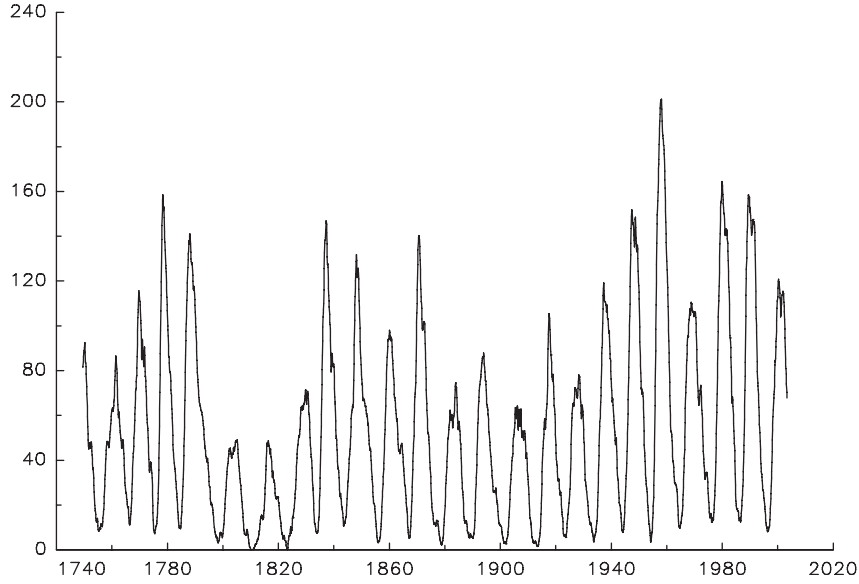


Fig. 8. Monthly Sunspot Numbers, 1749:07–2003:05.

with OLS:

$$\underset{}{\text{peak}} = \underset{(0.04)}{0.53} - \underset{(0.00079)}{0.0027}\,\text{mean}$$

$R^2 = 0.36$. However, the results seem to be heavily influenced by the observations from earlier sunspot cycles. For instance, Fig. 10 shows that earlier cycles have wider fluctuations in *peak* . Jagger (1977) contains a brief discussion on the quality of the sunspot numbers, and notes that the current definition of sunspot numbers dates only from 1849 and that the sunspot numbers are fairly subjective.[16] When only the recent 15 sunspot cycles are utilized, there appears to be little relation between the mean and peak of sunspot cycles as the following estimation result demonstrates:

$$\underset{}{\text{peak}} = \underset{(0.04)}{0.40} - \underset{(0.0007)}{0.0006}\,\text{mean}$$

$R^2 = 0.06$. Additionally, most of the *peak* values are less than 0.5, indicating asymmetry in the sunspot cycles, with the rise to maximum being faster than the fall to minimum. This asymmetry is well documented in the sunspot literature.

Table 2. Sunspot Cycles with Monthly Observations.

Cycle No.	Month of Minimum	Period (Months)	Mean
1			
2			
3			
4			
5			
6*	1755:03	135	41.7
7	1766:06	108	50.7
8	1775:06	111	67.1
9	1784:09	164	75.8
10	1798:05	151	28.4
11	1810:12	149	18.0
12	1823:05	126	20.6
13	1833:11	116	60.3
14	1843:07	149	64.3
15	1855:12	135	48.5
16	1867:03	141	61.7
17	1878:12	135	34.4
18	1890:03	143	41.2
19	1902:02	138	30.0
20	1913:08	120	33.4
21	1923:08	121	43.3
22	1933:09	125	47.8
23	1944:02	122	66.9
24	1954:04	126	91.9
25	1964:10	137	70.8
26	1976:03	126	67.8
27	1986:09	116	81.4
28**	1996:05	NA	NA

NA: Not applicable.
*Monthly observations are available from 1749:07 to 2003:05.
**The last cycle, Cycle 28, is not yet completed.

Finally, in Fig. 11, a scatter plot between the peak and length of sunspot cycle is displayed for the monthly observations.[17] G(57) noted that "The various points appear to lie on two distinct parabolas…" which suggests "two types of cycle, occurring randomly." However, it may be added that extreme observations in Fig. 11 are mainly from earlier sunspot cycles, which were potentially contaminated with measurement errors. With 15 recent sunspot cycles only, there appears to be little relation between the peak and length of sunspot cycles.

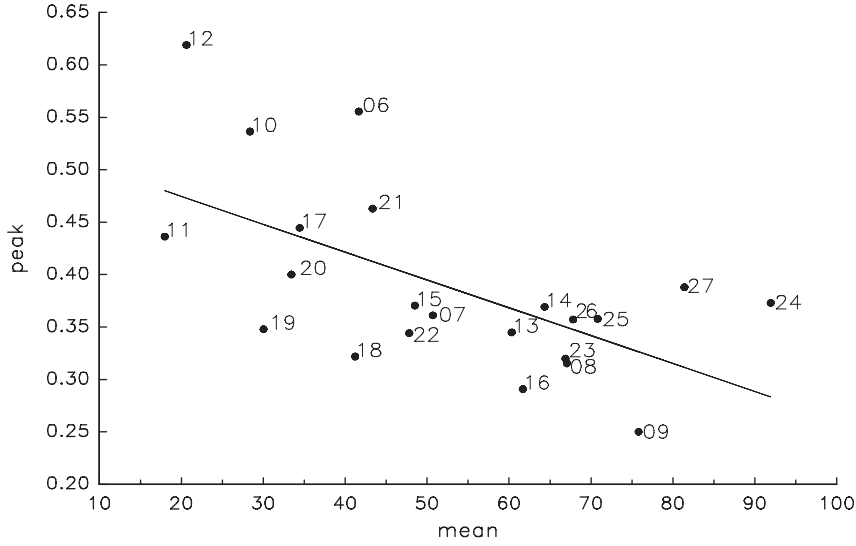


Fig. 9. Scatter Diagram Between Cycle Mean and Peak from Monthly Data.

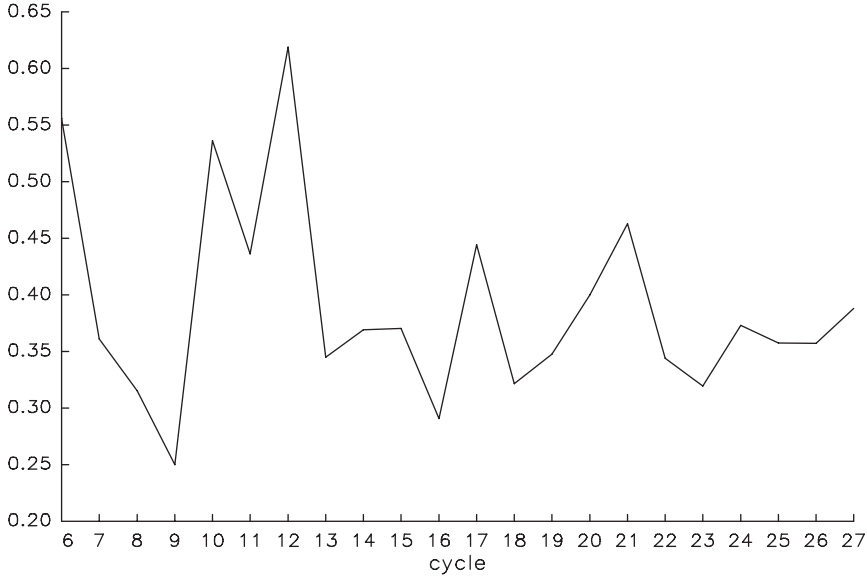


Fig. 10. Plot of Peak for Each Cycle from Monthly Data.

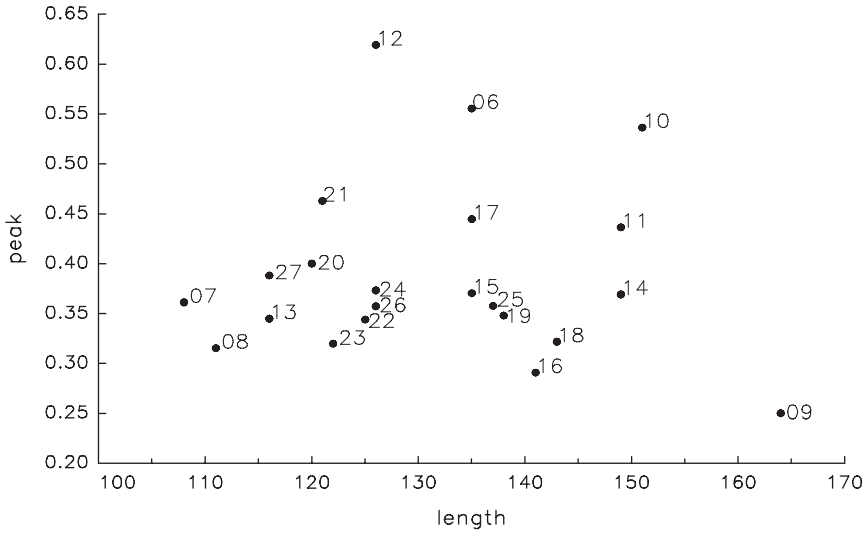


Fig. 11. Scatter Diagram Between Length of a Cycle and Peak from the Monthly Data, Corresponding to Fig. 3 in Granger (1957).

3.6. Forecasting Sunspot Numbers

G(57) offered no predictions for future sunspot values. Are these characterizations of sunspot cycles discussed in G(57) useful in forecasting the sunspot numbers? Because the two parameters in his model, the amplitude factor and the occurrence of minima, are hard to predict, G(57) noted that "It is thus doubtful that long-term prediction will be possible until a better understanding of the various solar variables is gained." However, once a cycle has begun, improved short-term sunspot forecasts would be possible. Forecasting of sunspot cycles is discussed in Morris (1977) and Hathaway et al. (1999), for instance. See also references cited therein.

4. CONCLUDING REMARKS

Clive Granger has permanently changed the shape of time series econometrics in his brilliant career spanning almost five decades. He has always been interested in problems empirically relevant to economic time series. In this paper, his very first published paper on sunspot numbers is reappraised

with extended sunspot numbers. The sunspot data are known to be very hard to analyze and forecast. His paper contained various interesting observations on sunspot cycles and is still cited in the literature. Whereas some of his findings appear to depend on particular earlier observations, which were possibly measured differently from later ones, most of his findings in his first published work still appear to hold with the extended sunspot numbers. His simple two-parameter mathematical model still accounts for more than 80% of the total variation in the extended sunspot data.

NOTES

1. The leading footnote of the paper reads that the paper was "Published with the aid of a grant from the American Astronomical Society."
2. The above-mentioned paper by Reichel et al. (2001) is unique in that, in addition to citing the first published work by Granger, it also tests for Granger-causality between the solar cycle length and the cycle mean of northern hemisphere land air temperature.
3. Granger was visiting Princeton in 1959–1960, holding a Commonwealth Fellowship of the Harkness Funds. He also spent three subsequent summers in Princeton.
4. To be precise, 27 sunspot cycles are used here from the extended sunspot numbers, compared to only 18 cycles in Granger (1957) for annual sunspot numbers. For monthly sunspot observations, four additional sunspot cycles are available in this paper. See Tables 1 and 2 below for more details.
5. This description of the deep meridional flow is from *the Economist* (2003).
6. For efficiency of exposition, further references to Granger (1957) will be designated as G(57).
7. Monthly sunspot numbers are also analyzed below. See Section 3 for more details.
8. The averaging was accomplished with arithmetic means. The yearly intensities in the quote denote the original sunspot numbers. See his Table 2 for more details.
9. The Figure corresponds to Fig. 1 in G(57).
10. Different functional forms are suggested for the basic curve in Elling and Schwentek (1992) and Hathaway et al. (1994).
11. The means are listed in the last column in Table 1, above. The Figure corresponds to Fig. 2 in G(57).
12. The expression in (3) for the long-term sunspot fluctuations is based on experiments, and no effort has been made to obtain an improved fit.
13. The notations in (4) are from G(57), in which the same argument x for both f and g was used. However, it appears that the notations should be more carefully expressed. The argument for function g should be the time from the minimum, as in (1), and the argument for function f changes for different sunspot cycles.
14. For ease of exposition, cycle numbers are also recorded in the Figure.

15. Smoothing is performed at the source with the 13-month running mean. If R_n is the sunspot number for month n, the 13-month running mean is $\bar{R}_n = \frac{1}{24}(\sum_{i=-6}^{5} R_{n+i} + \sum_{i=-5}^{6} R_{n+i})$.
16. Cycles post-1856 are understood to fall within what is known as the modern era of sunspot cycles.
17. The Figure corresponds to Fig. 3 in G(57).

ACKNOWLEDGMENTS

I would like to thank Sir Clive Granger, Tae-Hwan Kim, and Hee-Taik Chung for comments, and David Hendry and Jinki Kim for some references. An earlier version of this paper was presented at the "Predictive Methodology and Application in Economics and Finance" conference held in honor of Clive Granger at the University of California, San Diego, on January 6–7, 2004. The title of this paper was suggested to me by Norm Swanson. This research was supported by a Pusan National University Research Grant. Finally, some personal notes. I have had the privilege of working with Clive Granger for a long time, ever since I was a graduate student at UCSD. I gratefully appreciate his advice and encouragement over the years. I always learn a lot from him. On a cold winter night in London on December 6, 2002, he told me about his experience with his first publication. This paper is my small gift to him.

REFERENCES

Carozza, M., & Rampone, S. (2001). An incremental multivariate regression method for function approximation from noisy data. *Pattern Recognition*, *34*, 695–702.

Economist. (2003). *Sunspots: Staring at the Sun. Economist*, *367*(June 28), 116–117.

Elling, W., & Schwentek, H. (1992). Fitting the sunspot cycles 10–21 by a modified F-distribution density function. *Solar Physics*, *137*, 155–165.

Granger, C. W. J. (1957). A statistical model for sunspot activity. *Astrophysical Journal*, *126*, 152–158.

Granger, C. W. J. (1959). Estimating the probability of flooding on a tidal river. *Journal of the Institution of Water Engineers*, *13*, 165–174.

Granger, C. W. J. (1997). The *ET* interview, *Econometric Theory*, *13*, 253–303 (interviewed by P. C. B. Phillips).

Granger, C. W. J. (2001). Essays in econometrics. In: E. Ghysels, N. R. Swanson, & M. W. Watson (Eds), *Collected papers of Clive W. J. Granger* (Two Vols.), Econometric Society Monographs. Cambridge: Cambridge University Press.

Granger, C. W. J., & Andersen, A. P. (1978). *Introduction to bilinear time series models*. Göttingen: Vandenhoeck & Ruprecht.

Hathaway, D. H., Nandy, D., Wilson, R. M., & Reichman, E. J. (2003). Evidence that a deep meridional flow set the sunspot cycle period. *Astrophysical Journal*, *589*, 665–670.

Hathaway, D. H., Nandy, D., Wilson, R. M., & Reichman, E. J. (2004). Erratum: Evidence that a deep meridional flow set the sunspot cycle period. *Astrophysical Journal*, *602*, 543.

Hathaway, D. H., Wilson, R. M., & Reichman, E. J. (1994). The shape of the sunspot cycle. *Solar Physics*, *151*, 177–190.

Hathaway, D. H., Wilson, R. M., & Reichman, E. J. (1999). A synthesis of solar cycle prediction techniques. *Journal of Geophysical Research*, *104*, 22375–22388.

Hendry, D. F. (2004). The Nobel Memorial Prize for Clive W. J. Granger. *Scandinavian Journal of Economics*, *106*, 187–213.

Jagger, G. (1977). Discussion of M. J. Morris ''Forecasting the sunspot cycle''. *Journal of the Royal Statistical Society, Series A*, *140*, 454–455.

Morris, M. J. (1977). Forecasting the sunspot cycle. *Journal of the Royal Statistical Society, Series A*, *140*, 437–468.

Reichel, R., Thejll, P., & Lassen, K. (2001). The cause-and-effect relationship of solar cycle length and the Northern hemisphere air surface temperature. *Journal of Geophysical Research-Space Physics*, *106*, 15635–15641.

Shell, K. (2001). MD interview: An interview with Karl Shell. *Macroeconomic Dynamics*, *5*, 701–741 (interviewed by Stephen E. Spear and Randall Wright).

Stewart, J. Q., & Eggleston, F. C. (1940). The mathematical characteristics of sunspot variations II. *Astrophysical Journal*, *91*, 72–82.

Stewart, J. Q., & Panofsky, H. A. A. (1938). The mathematical characteristics of sunspot variations. *Astrophysical Journal*, *88*, 385–407.

Tong, H. (1990). *Non-linear time series*. A dynamic system approach, Oxford: Clarendon Press.

PERSONAL COMMENTS ON YOON'S DISCUSSION OF MY 1957 PAPER

Sir Clive W. J. Granger, KB

There is a type of art that is known as "naive," which can be very simplistic and have a certain amount of charm. Reading about my own work on sunspots published in 1957 also brings to mind the description "naive." I was certainly naive in thinking that I could undertake a piece of simple numerical analysis and then send it off to a major journal. The world of empirical research was then very simplistic as we had no computer at the University of Nottingham where I was employed, and all calculations had to be done on an electronic calculator and all plots by hand. It was clear that if monthly data for sunspots had been available, I would have been overwhelmed by the quantity of data! Trying to plot by hand nearly 200 years of monthly data is a lengthy task! The computing restrictions also limited the types of model that could be considered.

When I started the analysis I was employed as a statistician in the Department of Mathematics and so any data was relevant. I was intrigued by the sunspot series for several reasons. It was a long sequence with interesting features and seemed to be of interest to others. I also did not understand why it had been gathered. I was told it had been gathered by Swiss monks for a couple of hundred years, but no reason for their activity was given.

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 315–316
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20032-4

The obvious way to evaluate a model is to use an extended data set, as Gawon Yoon has done. However, more recent data are gathered in a different, more reliable manner, and so the error-rate on the series may have fallen. It is unclear to me how a ''sunspot'' is actually measured as it surely has indistinct edges and a different side of the sun will face us every month and year. I suppose that the 11-year cycle could be because some sides of the sun are more active than others, and the sun rotates every 11 years. I have absolutely no idea if this is correct; in fact, the analysis of sunspot numbers by statisticians is poor empirical analysis with no physical or solar theory used to form the model. How different this is than our econometrics models when economic theory provides some structure in many cases.

Not only do ''sunspots'' have an interesting place in economic theory thanks to the analysis of Cass and Shell (1983), among many, but they are also outliers in Physics. I would suppose Physics is now dominated by the idea that the world is deterministic (apart from the quantum domain) and so relies heavily on chaos theory and models. I surmise that those models cannot fit sunspot data – even the recent data – without substantial errors.

I think that returning to the occasional old analysis may be a worthwhile pursuit. I believe that my 1957 paper will prove to be not the best example available to provide lessons for current research.

I would like to thank Gawon Yoon for his paper and work.

REFERENCE

Shell, K., & Cass, D. (1983). Do sunspots matter? *Journal of Political Economy*, *91*, 193–227.

A NEW CLASS OF TAIL-DEPENDENT TIME-SERIES MODELS AND ITS APPLICATIONS IN FINANCIAL TIME SERIES

Zhengjun Zhang

ABSTRACT

In this paper, the gamma test is used to determine the order of lag-k tail dependence existing in financial time series. Using standardized return series, statistical evidences based on the test results show that jumps in returns are not transient. New time series models which combine a specific class of max-stable processes, Markov processes, and GARCH processes are proposed and used to model tail dependencies within asset returns. Estimators for parameters in the models are developed and proved to be consistent and asymptotically joint normal. These new models are tested on simulation examples and some real data, the S&P 500.

1. INTRODUCTION

Tail dependence – which is also known as asymptotic dependence or extreme dependence – exists in many applications, especially in financial

Econometric Analysis of Financial and Economic Time Series/Part B
Advances in Econometrics, Volume 20, 317–352
Copyright © 2006 by Elsevier Ltd.
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1016/S0731-9053(05)20033-6

time series analysis. Not taking this dependence into account may lead to misleading results. Towards a solution to the problem, we show statistical evidences of tail dependencies existing in jumps in returns from standardized asset returns, and propose new nonlinear time series models which can be used to model serially tail dependent observations.

A natural way of modeling tail dependencies is to apply extreme value theory. It is known that the limiting distributions of univariate and multivariate extremes are max-stable, as shown by Leadbetter, Lindgren and Rootzen (1983) in the univariate case and Resnick (1987) in the multivariate case. Max-stable processes, introduced by de Haan (1984), are an infinite-dimensional generalization of extreme value theory, and they do have the potential to describe clustering behavior and tail dependence.

Parametric models for max-stable processes have been considered since the 1980s. Deheuvels (1983) defines the moving minima (MM) process. Davis and Resnick (1989) study what they call the max-autoregressive moving average (MARMA) process of a stationary process. For prediction, see also Davis and Resnick (1993). Recently, Hall, Peng, and Yao (2002) discuss moving maxima models. In the study of the characterization and estimation of the multivariate extremal index, introduced by Nandagopalan (1990, 1994), Smith and Weissman (1996) extend Deheuvels' definition to the so-called multivariate maxima of moving maxima (henceforth M4) process.

Like other existing nonlinear time series models, motivations of using max-stable process models are still not very clear. First, real data applications of these models are yet becoming available. Also, statistical estimation of parameters in these models is not easy due to the nonlinearity of the models and the degeneracy of the joint distributions. A further problem is that a single form of moving maxima models is not realistic since the clustered blowups of the same pattern will appear an infinite number of times as shown in Zhang and Smith (2004). In our paper, we explore the practical motivations of applying certain classes of max-stable processes, GARCH processes and Markov processes.

In time series modeling, a key step is to determine a workable finite dimensional representation of the proposed model, i.e. to determine statistically how many parameters have to be included in the model. In linear time series models, such as AR(p), MA(q), auto-covariance functions or partial auto-covariance functions are often used to determine this dimension, such as the values of p and q. But in a nonlinear time-series model, these techniques may no longer be applicable. In the context of max-stable processes, since the underlying distribution has no finite variance, the

dimension of the model can not be determined in the usual manner. We use the gamma test and the concept of lag-k tail dependence, introduced by Zhang (2003a), to accomplish the task of model selection. Comparing with popular model selection procedures, for example Akaike's (1974) Information Criteria (AIC), Schwarz Criteria (BIC) (1978), our procedure first determines the dimension of the model prior to the fitting of the model.

In this paper, we introduce a base model which is a specific class of maxima of moving maxima processes (henceforth M3 processes) as our newly proposed nonlinear time-series model. This base model is mainly motivated through the concept of tail dependencies within time series. Although it is a special case of the M4 process, the motivation and the idea are new. This new model not only possesses the properties of M4 processes, which have been studied in detail in Zhang and Smith (2004), but statistical estimation turns out to be relatively easy.

Starting from the basic M3 process at lag-k, we further improve the model to allow for possible asymmetry between positive and negative returns. This is achieved by combining an M3 process with an independent two state ($\{0,1\}$) Markov chain. The M3 process handles the jumps, whereas the Markov chain models the sign changes. In a further model, we combine two independent M3 processes, one for positive and negative within each, with an independent three state Markov chain ($\{-1, 0, 1\}$) governing the sign changes in the returns.

In Section 2, we introduce the concepts of lag-k tail dependence and the gamma test. Some theoretical results are listed in that section; proofs are given in Section 9. In Section 4, we introduce our proposed model and the motivations behind it. The identifiability of the model is discussed. The tail dependence indexes are explicitly derived under our new model. In Section 5, we combine the models introduced in Section 4 with a Markov process for the signs. In Section 7, we apply our approach to the S&P 500 index. Concluding remarks are listed in Section 8. Finally, Section 9 continues the more technical proofs.

2. THE CONCEPTS OF LAG-K TAIL DEPENDENCE AND THE GAMMA TEST

Sibuya (1960) introduces tail independence between two random variables with identical marginal distributions. De Haan and Resnick (1977) extend it to the case of multivariate random variables. The definition of tail

independence and tail dependence between two random variables is given below.

Definition 2.1. A bivariate random variable (X_1, X_2) is called tail independent if

$$\lambda = \lim_{u \to x_F} P(X_1 > u | X_2 > u) = 0, \tag{2.1}$$

where X_1 and X_2 are identically distributed with $x_F = \sup\{x \in \mathbb{R} : P(X_1 \leq x) < 1\}$; λ is also called the bivariate tail dependence index which quantifies the amount of dependence of the bivariate upper tails. If $\lambda > 0$, then (X_1, X_2) is called tail dependent.

If the joint distribution between X_1 and X_2 is known, we may be able to derive the explicit formula for λ. For example, when X_1 and X_2 are normally distributed with correlation $\rho \in (0,1)$ then, $\lambda = 0$. When X and Y have a standard bivariate t-distribution with v degrees of freedom and correlation $\rho > -1$ then, $\lambda = 2\bar{t}_{v+1}(\sqrt{v+1}\sqrt{1-\rho}/\sqrt{1+\rho})$, where $\bar{t}_{v+1}$ is the tail of standard t distribution. Embrechts, McNeil, and Straumann (2002) give additional cases where the joint distributions are known.

Zhang (2003a) extends the definition of tail dependence between two random variables to lag-k tail dependence of a sequence of random variables with identical marginal distribution. The definition of lag-k tail dependence for a sequence of random variables is given below.

Definition 2.2. A sequence of sample $\{X_1, X_2, ..., X_n\}$ is called lag-k tail dependent if

$$\lambda_k = \lim_{u \to x_F} P(X_1 > u | X_{k+1} > u) > 0, \quad \lim_{u \to x_F} P(X_1 > u | X_{k+j} > u) = 0, j > 1, \tag{2.2}$$

where $x_F = \sup\{x \in \mathbb{R} : P(X_1 \leq x) < 1\}$; λ_k is called lag-k tail dependence index.

Here we need to answer two questions: the first is whether there is tail dependence; the second is how to characterize the tail dependence index. The first question can be put into the following testing of hypothesis problem between two random variables X_1 and X_2:

$$\begin{aligned} &H_0 : X_1 \quad \text{and} \quad X_2 \quad \text{are} \quad \text{tail} \quad \text{independent} \\ &\leftrightarrow H_1 : X_1 \text{ and } X_2 \text{ are tail dependent,} \end{aligned} \tag{2.3}$$

which can also be written as

$$H_0 : \lambda = 0 \quad vs. \quad H_1 : \lambda > 0. \tag{2.4}$$

To characterize and test tail dependence is a difficult exercise. The main problem for computing the value of the tail dependence index λ is that the distribution is unknown in general; at least some parameters are unknown.

Our modeling approach first starts with the gamma test for tail (in)dependence of Zhang (2003a) which also demonstrates that the gamma test efficiently detects tail (in)dependencies at high threshold levels for various examples. The test goes as follows.

Let

$$\begin{pmatrix} X_1, & X_2, & \ldots, & X_n \\ Y_1, & Y_2, & \ldots, & Y_n \end{pmatrix} \tag{2.5}$$

be an independent array of unit Fréchet random variables which have distribution function $F(x) = \exp(-1/x), x > 0$. Now let (U_i, Q_i), $i = 1, \ldots, n$ be a bivariate random sequence, where both U_i and Q_i are correlated and have support over (0,u] for a typically high threshold value u. Let $X_{ui} = X_i I_{\{X_i > u\}} + U_i I_{\{X_i \le u\}}$, $Y_{ui} = Y_i I_{\{Y_i > u\}} + Q_i I_{\{X_i \le u\}}$, $\quad i = 1, \ldots, n$. Then

$$\begin{pmatrix} X_{u1} \\ Y_{u1} \end{pmatrix}, \begin{pmatrix} X_{u2} \\ Y_{u2} \end{pmatrix}, \ldots, \begin{pmatrix} X_{un} \\ Y_{un} \end{pmatrix} \tag{2.6}$$

is a bivariate random sequence drawn from two dependent random variables X_{ui} and Y_{ui}. Notice that $X_{ui} I_{\{X_{ui} > u\}} (= X_i I_{\{X_i > u\}})$ and $Y_{ui} I_{\{Y_{ui} > u\}} (= Y_i I_{\{Y_i > u\}})$ are independent; but $X_{ui} I_{\{X_{ui} \le u\}} (= U_i I_{\{X_i \le u\}})$ and $Y_{ui} I_{\{Y_{ui} \le u\}} (= Q_i I_{\{Y_i \le u\}})$ are dependent. Consequently, if only tail values are concerned, we can assume the tail values are drawn from (2.5) under the null hypothesis of tail independence. We have the following theorem.

Theorem 2.3. Suppose V_i and W_i are exceedance values (above u) in (2.5). Then

$$P\left(\frac{u + W_i}{u + V_i} \le t\right) = \begin{cases} \frac{t}{1+t} - \frac{t}{1+t} e^{-(1+t)/u}, & \text{if} \quad 0 < t < 1 \\ \frac{t}{1+t} + \frac{t}{1+t} e^{-(1+t)/u}, & \text{if} \quad t \ge 1, \end{cases} \tag{2.7}$$

$$\lim_{n \to \infty} P(n^{-1}[\max_{i \le n}(u + W_i)/(u + V_i) + 1] \le x) = e^{-(1 - e^{-1/u})/x}. \tag{2.8}$$

Moreover,

$$\lim_{n \to \infty} P(n[\min_{i \le n}(u + W_i)/(u + V_i)] \le x) = 1 - e^{-(1 - e^{-1/u})x}. \tag{2.9}$$

The random variables $\max_{i\le n}\ (u+W_i)/(u+V_i)$ and $\max_{i\le n}\ (u+V_i)/(u+W_i)$ are tail independent, i.e.

$$\lim_{n\to\infty} P(n^{-1}[\max_{i\le n}\frac{(u+W_i)}{(u+V_i)}+1]\le x,\quad n^{-1}[\max_{i\le n}\frac{(u+V_i)}{(u+W_i)}+1]\le y)$$
$$= e^{-(1-e^{-1/u})/x-(1-e^{-1/u})/y}. \tag{2.10}$$

Furthermore, the random variable

$$Q_{u,n}=\frac{\max_{i\le n}\{(u+W_i)/(u+V_i)\}+\max_{i\le n}\{(u+V_i)/(u+W_i)\}-2}{\max_{i\le n}\{(u+W_i)/(u+V_i)\}\times\max_{i\le n}\{(u+V_i)/(u+W_i)\}-1} \tag{2.11}$$

is asymptotically gamma distributed, i.e. for $x\ge 0$,

$$\lim_{n\to\infty} P(Q_{u,n}\le x)=\zeta(x), \tag{2.12}$$

where ζ is gamma$(2, 1-e^{-1/u})$ distributed.

A proof of Theorem 2.3. is given in Zhang (2003a). Under certain mixing conditions, Zhang (2003a) also shows that (2.11) and (2.12) hold when V_i and W_i are not exceedances from (2.5)

Remark 1. If $V_i=(X_{ui}-u)I_{\{X_{ui}>u\}}$ and $W_i=(Y_{ui}-u)I_{\{Y_{ui}>u\}}$, then (2.11) and (2.12) hold.

Eqs. (2.11) and (2.12) together provide a gamma test statistic which can be used to determine whether tail dependence between two random variables is significant or not. When $nQ_{u,n}>\zeta_\alpha$, where ζ_α is the upper αth percentile of the gamma$(2, 1-e^{-1/u})$ distribution, we reject the null-hypothesis of no tail dependence.[1]

Remark 2. In the case of testing for lag-k tail dependence, we let $Y_i=X_{i+k}$ in the gamma test statistic (2.11). If the observed process is lag-j tail dependent, the gamma test would reject the H_0 of (2.3) when $k=j$, but it would retain the H_0 of (2.3) when $k=j+1$ and a bivariate subsequence $\{(X_{i_l}, Y_{i_l}),\ \ l=1, \ldots,\ \ i=1, \ldots\}$ of data are used to compute the value of the test statistic, where $i_l-i_{l-1}>j$. In Section 4, we will introduce a procedure using local windows to accomplish this testing procedure.

So far, we have established testing procedure to determine whether there exists tail dependence between random variables. Once the H_0 of (2.3) is rejected, we would like to model the data using a tail dependence model. In the next section we explore how the existence of tail dependence may imply transient behaviors in the jumps of financial time series.

3. JUMPS IN RETURNS ARE NOT TRANSIENT: EVIDENCES

The S&P 500 index has been analyzed over and over again, but models using extreme value theory to analyze S&P 500 index returns are still somehow delicate. There are several references; examples include Tsay (1999) and, Danielsson (2002). Models addressing tail dependencies within the observed processes are more difficult to find in the literature. In this section, we use the gamma test to check whether there exists tail dependence for the S&P 500 return data.

Recall that, in order to use the gamma test for tail independence, the original distribution function needs to be transformed into unit Fréchet distribution. This transformation preserves the tail dependence parameter.

3.1. Data Transformation and Standardization

In Fig. 1, we plot the negative log returns of the S&P500 index. The data are from July 3, 1962 to December 31, 2002. The highest value corresponds to

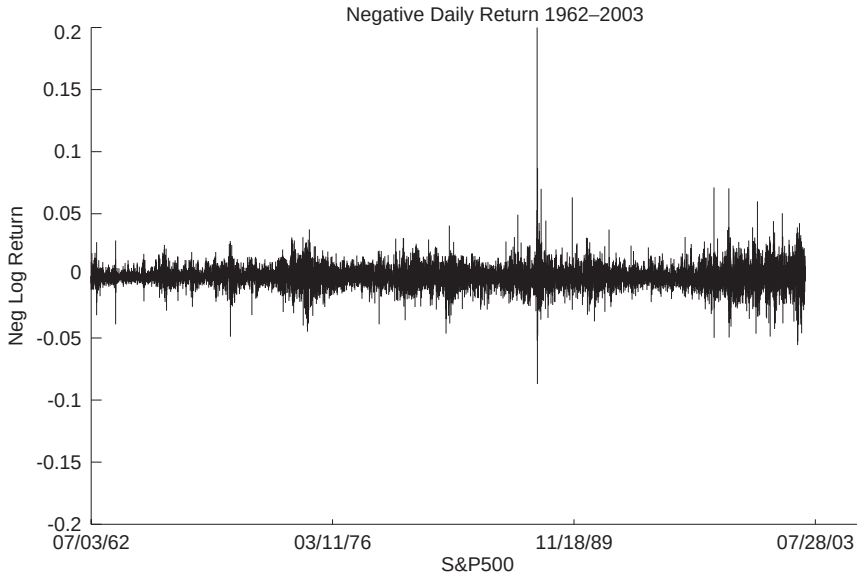


Fig. 1. Plot of the S&P 500 Original Log Returns. The Highest Value Corresponds to October 19, 1987 Wall Street Crash.

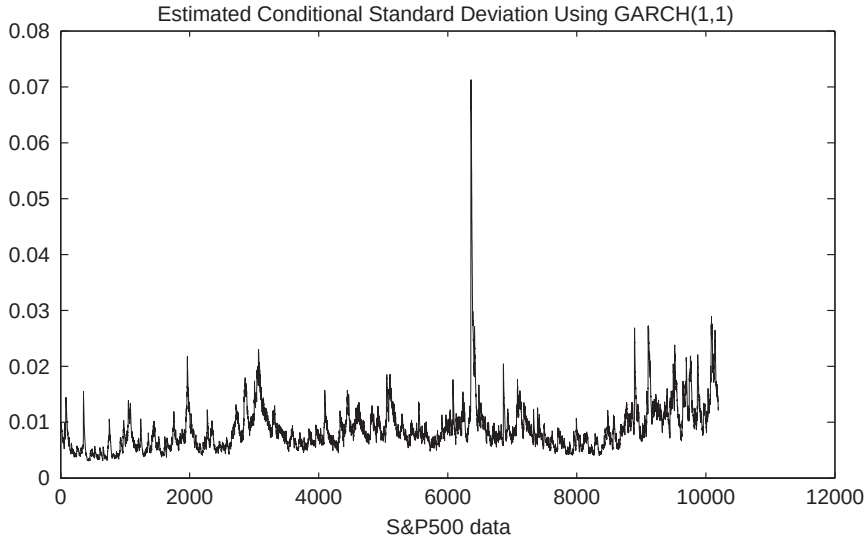


Fig. 2. Estimated Volatility Plot of the S&P 500 Log Returns. GARCH(1,1) Model is used to Estimate the Volatilities.

October 19, 1987 Wall Street crash. There clearly are large moves (possibly jumps) both in the returns as well as in the volatilities. We are especially interested in the tail dependence resulting from these large moves.

As GARCH models have been quite successful in modelling volatilities in financial time series, we apply a GARCH fitting to the data. A GARCH(1,1) (Bollerslev, 1986) model fitted to the data is plotted in Fig. 1. The estimated volatility process is plotted in Fig. 2. The original data, divided by the estimated standard volatility, are plotted in Fig. 3. This series is referred as the GARCH-devolatilized time series. Examples of papers going beyond this approach are for instance McNeil and Frey (2000), and Engle (2002). In this paper, we study the standardized financial time series, or GARCH residuals using our methods. We focus on the modeling of the larger jumps in the standardized returns.

Notice that the largest value in the standardized series does not correspond to October 19, 1987 Wall Street crash. The larger values in the standardized series look more like extreme observations, not outliers. An example of simulated extreme process which demonstrates extreme observations is illustrated in a later section. We now turn to study data transformation methods.

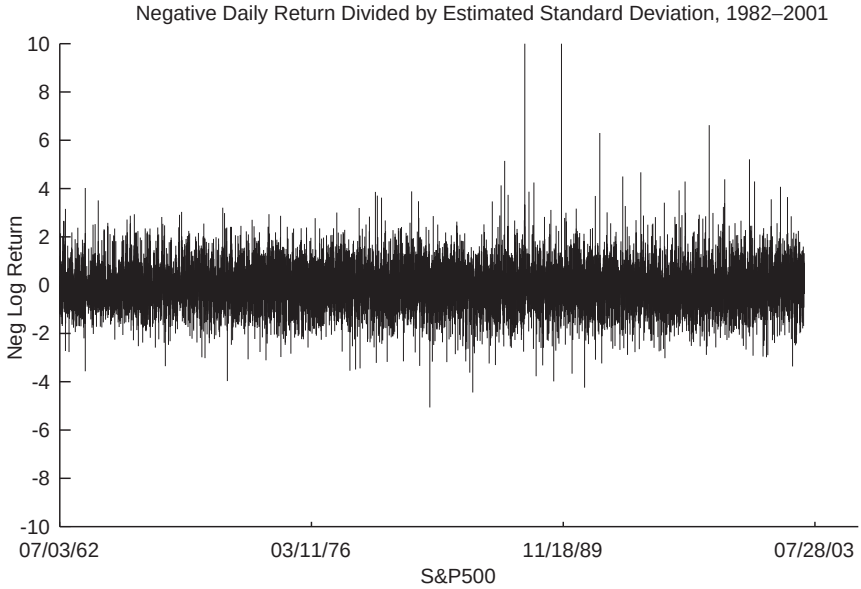


Fig. 3. Plot of the S&P 500 Standardized Log Returns.

Methods for exceedance modeling over high thresholds are widely used in the applications of extreme value theory. The theory used goes back to Pickands (1975), below we give a brief review. For full details, see Embrechts, Klüppelberg, and Mikosch (1997) and the references therein. Consider the distribution function of a random variable X conditionally on exceeding some high threshold u, we have

$$F_u(y) = P(X \le u + y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \quad y \ge 0.$$

As $u \to x_F = \sup\{x : F(x) < 1\}$, Pickands (1975) shows that the generalized Pareto distributions (GPD) are the only non-degenerate distribution that approximate $F_u(y)$ for u large. The limit distribution of $F_u(y)$ is given by

$$G(y; \sigma, \xi) = 1 - (1 + \xi \frac{y}{\sigma})_+^{-1/\xi}. \tag{3.1}$$

In the GPD, $y_+ = \max(y, 0)$, ξ is the tail index, also called shape parameter, which gives a precise characterization of the shape of the tail of the GPD. For the case of $\xi > 0$, it is long-tailed, i.e., $1 - G(y)$ decays at rate $y^{-1/\xi}$ for

large y. The case $\xi<0$ corresponds to a distribution function which has a finite upper end point at $-\sigma/\xi$. The case $\xi=0$ yields the exponential distribution with mean σ:

$$G(y;\sigma,0)=1-\exp(-\frac{y}{\sigma}).$$

The Pareto, or GPD, and other similar distributions have long been used models for long-tailed processes. As explained in the previous sections, we first transform the data to unit Fréchet margins. This is done using the generalized extreme value (GEV) distributions which are closely related to the GPD as explained in Embrechts et al. (1997) and can be written as

$$H(x)=\exp\left[-(1+\xi\frac{x-\mu}{\psi})_{+}^{-1/\xi}\right], \tag{3.2}$$

where μ is a location parameter, $\psi>0$ is a scale parameter, and ξ is a shape parameter similar to the GPD form in (3.1). Pickands (1975) first established the rigorous connection of the GPD with the GEV. Using the GEV and GPD, we fit the tails of the negative observations, the positive observations, and the absolute observations separately. The absolute returns are included in our data analysis for comparison. The generalized extreme value model (3.2) is fitted to observations which are above a threshold u. We have tried a series of threshold values and performed graphical diagnosis using the mean excess plot, the W and Z statistics plots due to Smith (2003). Those diagnosis suggested that when values above $u=1.2$, visually, a GPD function fits data well. There are about 10% of observed values above the threshold $u=1.2$. The maximum likelihood estimates are summarized in Table 1. From the table, we see that standardized negative returns are still fat tailed since the estimated shape parameter value is positive, but the

Table 1. Estimations of Parameters in GEV Using Standardized Return Series and Threshold Value 1.2. N_u is the Number of Observations Over Threshold u.

Series	N_u	μ (SE)	Logψ (SE)	ξ (SE)
Negative	1050	3.231845	−0.319098	0.096760
		(0.086479)	(0.075502)	(0.028312)
Positive	1088	2.869327	−0.784060	−0.062123
		(0.054016)	(0.064046)	(0.025882)
Absolute	2138	3.514718	−0.450730	0.045448
		(0.071042)	(0.060214)	(0.018304)

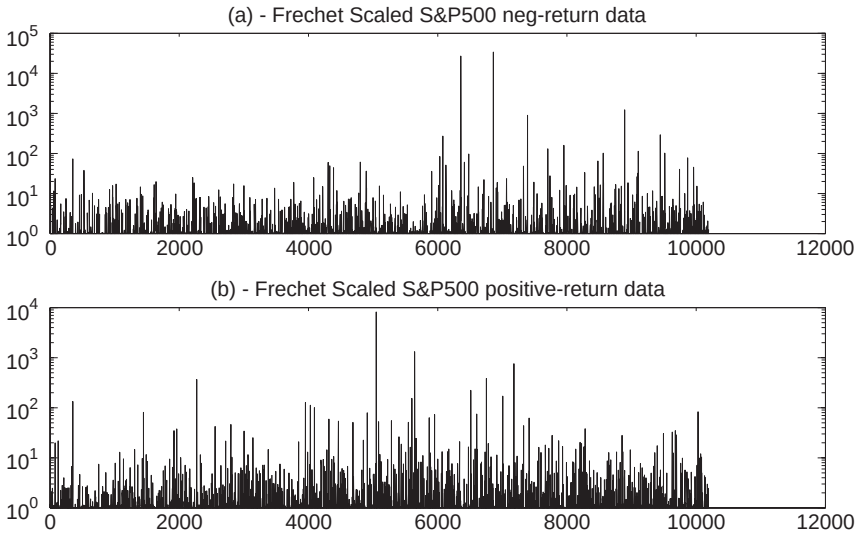


Fig. 4. Plots of the S&P 500 Standardized Time Series in Fréchet Scale. Panel (a) is for Negative Returns. Panel (b) is for Positive Returns. The Threshold used for Both Plots is $u = 1.2$.

standardized positive returns are now short tailed. The absolute returns are still fat tailed.

The standardized data are now converted into unit Fréchet scale for the observations above $u = 1.2$ using the tail fits resulting from Table 1. The final transformed data are plotted in Fig. 4. After standardization and scale transformation, we can say that the data show much less jumps in volatility, but jumps in returns are still persistent as will be shown in the next Section.

3.2. Tail Dependence

In the applications of GARCH modelling, GARCH residuals are assumed to be independently identically distributed normal (or *t)* random variables. The residual distribution assumptions may not be appropriate in some applications. As a result, there have been various revised GARCH type models proposed in the literature. Hall and Yao (2003), Mikosch and Stărică (2000), Mikosch and Straumann (2002) – among others – have discussed heavy-tailed errors and extreme values. Straumann (2003) deals with the estimation in certain conditionally heteroscedastic time series models, such as GARCH(1,1), AGARCH(1,1), EGARCH etc. Our main

interest is the analysis of tail dependence within the standardized unit Fréchet transformed returns. Given such tail dependence, we want to present a relevant time series model capturing that behavior.

In order to check whether there exists tail dependence within the sequences in Fig. 4, we perform the gamma test from Section 2 to the transformed returns with the threshold level at the 95th percentile (90th percentile is related to $u = 1.2$) of observed sequence as in Zhang (2003a).

However, in reality or even in simulated examples as shown in Section 4, one single test may not be enough to discover the true tail dependence or independence. As pointed out in Zhang (2003a), the gamma test is able to detect tail dependence between two random variables even if there are outliers in one of two observed sequences. In our case, we have a univariate sequence. If there are outliers in the sequence, a Type I error may occur since outliers cause smaller $Q_{w,n}$ values in (2.11). If there are lower lag tail dependencies, a Type II error may occur since there are dependencies in each projected sequence. In order to minimize the probability of making errors, we select observations in local windows of size 500 to perform the gamma test. The observations in each window are used to compute the value of the gamma test statistic and a global conclusion is suggested at level $\alpha = 0.05$. Besides the gamma test, we also compute the empirical estimation of the tail dependence indexes. The empirical estimation is done by using the following procedure.

Empirical estimation procedures: Suppose $x_1, x_2, \ldots, x_n$ is a sequence of observed values and x^* is the 95th sample percentile. Then the empirical estimate of lag-k tail dependence index is $\sum_{i=1}^{n-k} I_{(x_i > x^*, x_{i+k} > x^*)} \Big/ \sum_{i=1}^{n-k} I_{(x_i > x^*)}$, where $I_{(.)}$ is an indictor function. The test results for lags $1-15$ are summarized in Table 2.

Columns 2, 6 ($\times$ l00) yield percentages of rejection of H_0 from all fully enumerated local windows of size 500. Of each data point, at least one component is nonzero. Columns 3, 7 are empirically estimated lag-k tail dependence indexes over a threshold value computed at the 95th percentile for the whole data. Columns 4, 8 are the minima of all computed $Q_{w,n}$ values using (2.11) in all local windows. Columns 5, 9 are the maxima of all computed $Q_{w,n}$ values using (2.11). The number 0.7015 in Column 2 means that if we partition the whole data set into 100 subsets, and perform the gamma test to each subset, there are about 70 subsets from which we would reject the tail independence hypothesis. The number 0.0700 in Column 3 is the empirically estimated tail dependence index which tells that when a large price drop is observed, and the resulting negative return is below the 95th percentile of the historical data, there is 7% chance to observe a large price drop in the kth day. The rest of numbers in the table can be interpreted similarly.

Table 2. Columns 2, 6 ($\times$ l00) Yield Percentages of Rejection of H_0 from all Fully Enumerated Local Windows of Size 500. Of Each Data Point, at Least One Component is Nonzero. Columns 3, 7 are Estimated Lag-k Tail Dependence Indexes Over a Threshold Value Computed at the 95th Percentile for the Whole Data. Columns 4, 8 are the Minima of all Computed $Q_{u,n}$ Values using (2.11) in all Local Windows. Columns 5, 9 are the Maxima of all Computed $Q_{u,n}$ Values Using (2.11).

lag	Negative Returns				Positive Returns			
k =	Rej.	Ind.	Min	Max	Rej.	Ind.	Min	Max
1	1.0000	0.0700	0.0162	0.3092	0.6787	0.0555	0.0079	0.1291
2	0.7015	0.0700	0.0013	0.2292	0.7556	0.0357	0.0078	0.1905
3	0.6064	0.0525	0.0125	0.2370	0.6454	0.0317	0.0126	0.1630
4	0.8440	0.0525	0.0185	0.4046	0.4104	0.0317	0.0152	0.1170
5	0.5713	0.0306	0.0095	0.3369	1.0000	0.0238	0.0344	0.2113
6	0.4452	0.0088	0.0174	0.2761	0.5147	0.0119	0.0181	0.1597
7	1.0000	0.0131	0.0532	0.2917	0.5125	0.0159	0.0196	0.2245
8	0.7417	0.0306	0.0029	0.1353	0.1158	0.0079	0.0067	0.0485
9	0.3793	0.0175	0.0244	0.3740	0	0.0119	0.0065	0.0268
10	1.0000	0.0131	0.0264	0.0687	0	0.0159	0.0060	0.0186
11	1.0000	0.0088	0.0262	0.2012	1.0000	0.0159	0.0309	0.0444
12	1.0000	0.0131	0.0556	0.1802	1.0000	0.0079	0.0530	0.0679
13	0	0.0175	0.0083	0.0093	0.8681	0.0040	0.0219	0.0320
14	1.0000	0.0088	0.0480	0.0528	0	0.0119	0.0072	0.0081
15	0.5217	0.0131	0.0098	0.0436	1.0000	0.0198	0.0839	0.1061

From Column 2 in Table 2, we see that for lags from 1 to 12, the percentages of rejecting H_0 are high for negative returns. At lags 1, 7, 10, 11, 12, tail independence is always rejected in the case of negative returns. In Column 3, if adding up all estimated lag-1 to lag-12 tail dependence indexes, the total percentage is over 30% which tells that there is more than 30% of chance that a large price drop may happen in one of the following 12 days. This should be seriously taken into account in modeling. Lag-13 is a break point at which the rejection rate of tail independence is 0 for negative returns. The empirically estimated dependence indexes in Column 3 show a decreasing trend at the beginning five lagged days; then after lag-5, they become small and no trend. The table also shows that each of the first five empirically estimated tail dependence indexes is falling in the range of the minima of $Q_{u,n}$ value, and the maxima of $Q_{u,n}$. These numbers suggest that there is tail dependence within the negative returns. All the empirically computed values, the $Q_{u,n}$ values can be used as estimates of the tail

dependence index. In Section 4, we compute theoretical lag-k tail dependence indexes for certain models.

Columns 6–9 are for positive returns, and can be interpreted in a similar way to negative returns. Column 6 shows that up to lag-8, the positive returns are tail dependent. Lag-5 tail independence is always rejected. Lag-9 is a break point at which the rejection rate of tail independence is 0. Like negative returns, Column 7 shows a decreasing trend of the empirically estimated tail dependence indexes. After lag-5, there is no pattern. Columns 7,8 are the minima and the maxima of empirical tail dependence index values from formula (2.11). Each of the first five numbers in Column 7 is falling in the range of the minima and the maxima in Columns 8, 9 with the same lag numbers.

In the literature, that jumps in returns are transient often refers to that a large jump (positive or negative) does not have a future impact. When jumps in return series are transient, they are tail independent. Models dealing with transience of jumps in return series are for instance Eraker, Johannes, and Polson (2003), Duan, Ritchken, and Sun (2003). Our results (jumps in returns being tail dependent) suggest that jumps in returns are not transient. When jumps in returns are tail dependent, a large jump may have an extreme impact on the future returns, and should be taken into account in modeling.

In order to find a criterion for deciding on the maximal lag of which tail dependence exists, we propose the following procedure:

Empirical criterion: When the rejection rate of the lag-r test first reaches 0 or a rate substantially smaller than the rejection rate of the lag-$(r-1)$ test, and the estimated lag-r tail dependence index is less than 1%, let $k = r$.

At the moment, this test procedure is still purely *ad hoc;* more work on its statistical properties is needed in the future. Based on this criteria, however, we use lag-12 tail dependence for the transformed negative returns and lag-7 tail dependence for the transformed positive returns in the sections below.

The above test may suggest the existence of tail dependence; finding on appropriate time series model explaining these dependencies is another task. In the next section we introduce a versatile class of models aimed at achieving just that.

4. THE BASIC MODEL AND THE COMPUTATION OF LAG-*K* TAIL DEPENDENCE INDEXES

Suppose the maximal lag of tail dependence within a univariate sequence is K, and $\{Z_{li}, l = 1, ..., L, -\infty < i < \infty\}$ is an independent array, where the

random variables Z_{li} are identically distributed with a unit Fréchet distribution function. Consider the model:

$$Y_i = \max_{0\le l\le L} \max_{0\le k\le K} a_{lk} Z_{l,i-k}, \quad -\infty < i < \infty, \tag{4.1}$$

where the constants $\{a_{lk}\}$ are nonnegative and satisfy $\sum_{l=0}^{L}\sum_{k=0}^{K} a_{lk} = 1$. When $L = 0$, model (4.1) is a moving maxima process. It is clear that the condition $\sum_{l=0}^{L}\sum_{k=0}^{K} a_{lk} = 1$ makes the random variable Y_i a unit Fréchet random variable. So Y_i can be thought of as being represented by an independent array of random variables which are also unit Fréchet distributed. The main idea behind (4.1) is that stochastic processes (Y_i) with unit Fréchet margins can be thought of as being represented (through (4.1)) by an infinite array of independent unit Fréchet random variables. A next step concerns the quality of such an approximation for finite (possibly small) values of L and K. The idea and the approximation theory are presented in Smith and Weissman (1996), Smith (2003). Zhang (2004) establishes conditions for using a finite moving range M4 process to approximate an infinite moving range M4 process.

Under model (4.1), when an extreme event occurs or when a large Z_{li} occurs, $Y_i \propto a_{l,i-k}$ for $i \approx k$, i.e. if some Z_{lk} is much larger than all neighboring Z values, we will have $Y_i = a_{l,i-k} Z_{lk}$ for i near k. This indicates a moving pattern of the time series, known as signature pattern. Hence, L corresponds to the maximal number of signature patterns. The constant K characterizes the range of dependence in each sequence and the order of moving maxima processes. We illustrate these phenomena for the case of $L = 0$ in Fig. 5. Plots (b) and (c) involve the same values of $(a_{00}, a_{01} \ldots, a_{0k})$. Plot (b) is a blowup of a few observations of the process in (a) and Plot (c) is a similar blowup of a few other observations of the process in (a). The vertical coordinate scales of Y in Plot (b) are from 0 to 5, while the vertical coordinate scales of Y in Plot (c) are from 0 to 50. These plots show that there are characteristic shapes around the local maxima that replicate themselves. Those blowups, or replicates, are known as signature patterns.

Remark 3. Model (4.1) is a simplified model of a general M4 process since we restrict our attention to univariate time series. In Zhang and Smith (2003, 2004), properties of models with a finite number of parameters are discussed. They provide sufficient conditions from which the estimators of model parameters are shown to be consistent and jointly asymptotic multivariate normal.

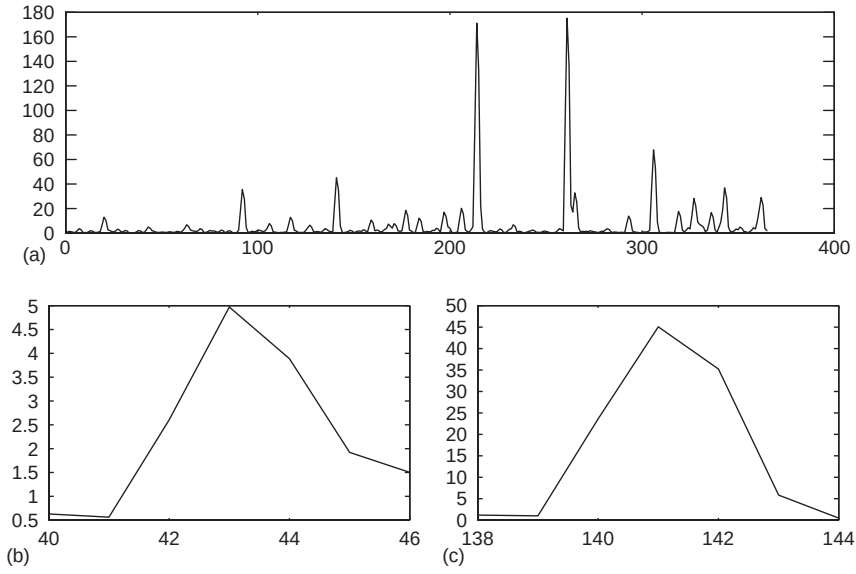


Fig. 5. An Illustration of a *M4* Process. Figure (a) consists of 365 Simulated Day Observation. Figures. (b), (c) are Partial Pictures Drawn from (a), i.e., (b) Is the Enlarged View of (a) From the 40th Day to the 45th Day; (c) is the enlarged view of (a) From the 138th Day to the 144th Day. (b) and (c) are Showing a Single Moving Pattern, called Signature Pattern, in Certain Time Periods when External Events Occur.

Under model (4.1), we have the following lag-k tail dependence index formula:

$$\lambda_k = \sum_{l=1}^{\infty} \sum_{m=-\infty}^{\infty} \min(a_{l,1-m}, a_{l,1+k-m}). \tag{4.2}$$

Obviously, as long as both a_{l0} and a_{lK} are non-zero, Y_i and Y_{i+K} are dependent, and of course tail dependent as can be seen from (4.2).

In real data, for example, the negative returns, we count how many times that negative return values $Y_i, Y_{i+1}, ..., Y_{i+k}, (i = 1, ..., n-k)$, are simultaneously greater than a given threshold value for any fixed k. We record all the i values such that $Y_i, Y_{i+1}, ..., Y_{i+k}$ are simultaneously greater than the given threshold value . We typically find that jumps in negative returns as

well as jumps in positive returns appear to cluster in two consecutive days, i.e., $k = 1$. In general, it is not realistic to observe similar blowup patterns of a period of more than two days. A preliminary data analysis shows that blowups seem to appear in a short time period (2 or 3 days), while the transformed returns (negative, positive) have a much longer lag-k tail dependence.

Considering the properties of clustered data in two consecutive days and much longer lag-k tail dependencies in real data, the following parameter structure seems reasonable, i.e., we assume the matrix of weights (a_{lk}) to have the following structure:

$$(a_{lk}) = \begin{pmatrix} a_{00} & 0 & 0 & 0 & \dots & 0 \\ a_{10} & a_{11} & 0 & 0 & \dots & 0 \\ a_{20} & 0 & a_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ a_{L0} & 0 & 0 & 0 & \dots & a_{LL} \end{pmatrix}. \tag{4.3}$$

Now, the number L corresponds to the maximal lag of tail dependencies within the sequence; the lag-k tail dependence index is characterized by the coefficients a_{k0} and a_{kk}. The coefficient a_{00} represents the proportion of the number of observations which are drawn from an independent process $\{Z_{0i}\}$. In other words, a very large value at time 0 has no future impact when the large value is generated from $\{Z_{0i}\}$. If both a_{k0} and a_{kk} are not zero, then a very large value at time 0 has impact at time k when the large value is generated from $\{Z_{ki}\}$. If there is strong lag-k tail dependence for each k, the value of a_{00} will be small. While two coefficients a_{k0} and a_{kk} may not be sufficient enough to characterize different kinds of k step impacts from a very large value at time 0, the setups of (4.1) and (4.3) are considered as an approximation of the observed process.

It is clear that as long as the maximal lag of tail dependence of a sequence has been determined, we can easily write the model based on (4.1) and (4.3). Although the estimators for coefficients in a general M4 model have been proposed by Zhang and Smith (2003), Zhang (2003b), there would be advantages to use the special structure in (4.3) to construct estimators, i.e., the conditions imposed on the parameters can be reduced to a minimal level. We now compute the lag-r tail dependence index and then, in next section, turn to the estimation of the parameters in (4.3).

For $r>0$, we have

$$\begin{aligned}
&P(Y_1 \le x, Y_{1+r} \le y) \\
&\quad = P(a_{lk} Z_{l,1-k} \le x,\ a_{lk} Z_{l,1+r-k} \le y,\ 0 \le l \le L,\ 0 \le k \le L) \\
&\quad = P(a_{l0} Z_{l1} \le x,\ a_{ll} Z_{l,1-l} \le x,\ a_{l0} Z_{l,1+r} \le y,\ a_{ll} Z_{l,1+r-l} \le y, \\
&\qquad 0 \le l \le L) \\
&\quad = \exp\left\{ -\sum_{l \ne r} \left(\frac{a_{l0}+a_{ll}}{x} + \frac{a_{r0}+a_{rr}}{y}\right) - \frac{a_{rr}}{x} - \frac{a_{r0}}{y} - \max\left(\frac{a_{r0}}{x}, \frac{a_{rr}}{y}\right) \right\} \\
&\quad = \exp\left\{ -\frac{1}{x} - \frac{1}{y} + \min\left(\frac{a_{r0}}{x}, \frac{a_{rr}}{y}\right) \right\}. \qquad (4.4)
\end{aligned}$$

A general joint probability computation leads to the following expression:

$$\begin{aligned}
P(Y_i \le y_i, 1 \le i \le r) &= P\left(Z_{l,i-k} \le \frac{y_i}{a_{l,k}} \quad \text{for} \quad 0 \le l \le L,\ 0 \le k \le L,\ 1 \le i \le r\right) \\
&= P\left(Z_{l,m} \le \min_{1-m \le k \le r-m} \frac{y_{m+k}}{a_{l,k}}, 0 \le l \le L, -l+1 < m < r\right) \\
&= \exp\left(-\sum_{l=0}^{L} \sum_{m=-l+1}^{r} \max_{1-m \le k \le r-m} \frac{a_{l,k}}{y_{m+k}} \right). \qquad (4.5)
\end{aligned}$$

Since

$$\begin{aligned}
\frac{P(Y_1 \ge u, Y_{1+r} \ge u)}{P(Y_1 \ge u)} &= 1 - \frac{\exp\{\frac{1}{u}\} - \exp\{-\frac{2-\min(a_{r0}, a_{rr})}{u}\}}{1 - \exp\{\frac{1}{u}\}} \\
&\to \min(a_{r0}, a_{rr}) \qquad (4.6)
\end{aligned}$$

as $u \to \infty$, the lag-r tail dependence index of model (4.1) is $\min(a_{r0}, a_{rr})$. There is some intuition behind the characterization of lag-r tail dependence index in (4.6). The value of $(a_{r0}+a_{rr})$ represents the proportion of the number of observations which are drawn from the process Z_{ri}. The value of $\min(a_{r0}, a_{rr})$ represents the proportion of the number of observations which are over a certain threshold and are drawn from the lag-r dependence process. This can be seen from the left hand side of (4.6) when it is replaced by its empirical counterpart. The empirical counterpart can also be used to estimate $\min(a_{r0}, a_{rr})$ for a suitable choice of u value. Parameter estimation will be discussed in Section 6. We now illustrate an example and show how the gamma test detects the order of lag-k tail dependence.

Example 4.1. Consider model (4.1) with the following parameter structure:

$$(a_{lk}) = \begin{pmatrix} 0.3765 & 0 & 0 & 0 & 0 & 0 \\ 0.0681 & 0.0725 & 0 & 0 & 0 & 0 \\ 0.0450 & 0 & 0.0544 & 0 & 0 & 0 \\ 0.0276 & 0 & 0 & 0.1166 & 0 & 0 \\ 0.0711 & 0 & 0 & 0 & 0.0185 & 0 \\ 0.1113 & 0 & 0 & 0 & 0 & 0.0386 \end{pmatrix}. \quad (4.7)$$

We use (4.1) and (4.7) to generate a sequence of observations of size 5000. These observations are plotted in Fig. 6. We use the gamma test (2.12) to test lag-k tail depenencies at level $\alpha = 0.05$.

As pointed out earlier, the index l in model (4.1) corresponds to a signature pattern of the observed process. The sum of $\sum_k a_{lk}$ yields the proportion of the total number of observations drawn from the lth independent sequences of Z_{lk} ,$-\infty < k < \infty$. If we simply use all data for the gamma test, we may not get the right indications of the lag-k tail dependencies because the values computed from the test statistic may not be associated with the kth moving

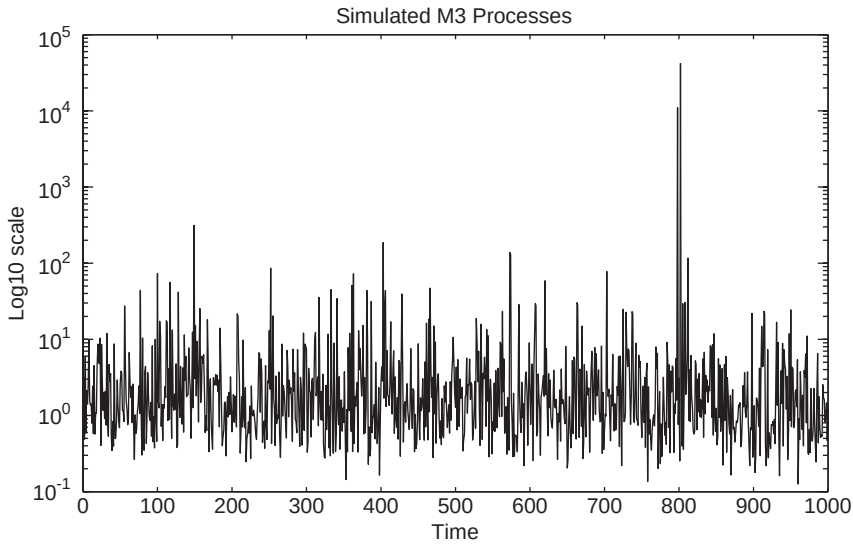


Fig. 6. Simulated Time Series of M3 Processes in Model (4.1).

pattern. Here we conduct the gamma test in a local moving window of size 300. We randomly draw 100 local windows and perform the gamma test using the 300 observations in each window. The number of rejections of lag-k tail dependencies are summarized in the following table.

lag-k	1	2	3	4	5	6	7	8	9	10
# of rejection of H_0	7	5	2	3	5	1	1	0	0	0

The rejection rates of the lag-k, $k = 1, 2, 3, 4, 5$, are close to their corresponding proportions of observations which are drawn from the corresponding moving patterns. The rejection rates of larger lag-k are relatively small. Therefore the maximal lag of 5 for tail dependence present in the simulated data is indeed suggested by the gamma test results.

5. COMBINING M3 WITH A MARKOV PROCESS: A NEW NONLINEAR TIME-SERIES MODEL

The previous analysis has suggested that there is asymmetric behavior between negative returns and positive returns. We have also tested whether there is tail dependence between positive returns and negative returns, and found that the null hypothesis of tail independence was not rejected. These phenomena suggest that models for negative returns should be different from models for positive returns.

First, we want to find a class of models of (4.1) and (4.3) to model negative returns. Second, we want to find a different class of models of (4.1) and (4.3) to model positive returns. Then we combine those two classes of models with a Markov process for both returns.

Notice that in the time series plot of negative returns, we have many zeros values (close to 50% of the total number of points) at which positive returns were observed. This suggests that models for negative returns should have a variable to model the locations of occurrences of zeros. This results in the following model structure.

Model 1. : Combining M3 (used to model scales) with a Markov process (used to model signs): a new model for negative returns:

$$Y_i^- = \max_{0\le l\le L^-} \max_{0\le k\le K^-} a_{lk}^- Z_{l,i-k}^-, \quad -\infty < i < \infty,$$

where the superscript $^-$ means that the model is for negative returns only. Constants $\{a_{lk}^-\}$ are nonnegative and satisfy $\sum_{l=0}^{L^-}\sum_{k=0}^{K^-} a_{lk}^- = 1$. The matrix of weights is

$$(a_{lk}^-) = \begin{pmatrix} a_{00}^- & 0 & 0 & 0 & \dots & 0 \\ a_{10}^- & a_{11}^- & 0 & 0 & \dots & 0 \\ a_{20}^- & 0 & a_{22}^- & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ a_{L^-0}^- & 0 & 0 & 0 & \dots & a_{L^-L^-}^- \end{pmatrix}.$$

$\{Z_{li}^-,\ l = 1, \dots, \mathrm{L}^-, -\infty < i < \infty\}$ is an independent array, where random variables Z_{li}^- are identically distributed with a unit Fréchet distribution function. Let

$$R_i^- = \xi_i^- Y_i^-, -\infty < i < \infty, \tag{5.1}$$

where the process $\{\xi_i^-\}$ is independent of $\{Y_i^-\}$ and takes values in a finite set $\{0, 1\}$ –i.e., $\{\xi_i^-\}$ is a sign process. Here $\{Y_i^-\}$ is an M3 process, $\{\xi_i^-\}$ is a simple Markov process. $\{R_i^-\}$ is the negative return process. For simplicity, Model (5.1) is regarded as MCM3 processes.

Remark 4. If $\{Y_i^-\}$ is an independent process, then $P(R_{i+r}^- > u | R_i^- > u) \to 0$ as $u \to \infty$ for $i > 0$, $\mathrm{r} > 0$, i.e., no tail dependence exists. This phenomenon tells that if there are tail dependencies in the observed process, the model with time dependence (through a Markov chain) only can not model the tail dependence if the random variables used to model scales are not tail dependent.

Model 2. : An MCM3 process model for positive returns:

$$Y_i^+ = \max_{0 \le l \le L^+} \max_{0 \le k \le K^+} a_{lk}^+ Z_{l,i-k}^+, -\infty < i < \infty,$$

where the superscript $^+$ means that the model is for positive returns only. Constants $\{a_{lk}^+\}$ are nonnegative and satisfy $\sum_{l=0}^{L^+}\sum_{k=0}^{K^+} a_{lk}^+ = 1$. The

matrix of weights is

$$(a_{lk}^{+}) = \begin{pmatrix} a_{00}^{+} & 0 & 0 & 0 & \dots & 0 \\ a_{10}^{+} & a_{11}^{+} & 0 & 0 & \dots & 0 \\ a_{20}^{+} & 0 & a_{22}^{+} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ a_{L^{+}0}^{+} & 0 & 0 & 0 & \dots & a_{L^{+}L^{+}}^{+} \end{pmatrix}.$$

Random variables $\{Z_{li}^{+},\ l = 1, \dots, \mathrm{L}^{+}, -\infty < i < \infty\}$ is an independent array, where Z_{li}^{+}s identically distributed with a unit Fréchet distribution. Let

$$R_i^{+} = \xi_i^{+} Y_i^{+}, \quad -\infty < i < \infty, \tag{5.2}$$

where the process $\{\xi_i^{+}\}$ is independent of $\{Y_i^{+}\}$ and takes values in a finite set $\{0, 1\}$. Here $\{Y_i^{+}\}$ is an M3 process, $\{\xi_i^{+}\}$ is a simple Markov process. $\{R_i^{+}\}$ is the positive return process.

Remark 5. In previous sections, we have seen that negative returns Y_i^{-} and positive returns Y_i^{+} are asymmetric, and concluded that models for positive returns should be different from models for negative returns. Notice that at any time i, one can only observe one of the Y_i^{-} and Y_i^{+}. The other one is missing. By introducing the Markov processes ξ_i^{-} and ξ_i^{+}, both R_i^{-} in (5.1) and R_i^{+} in (5.2) are observable. We use R_i^{-} and R_i^{+} to construct parameter estimators.

Model 3. : An MCM3 process model for returns: with the established notations in (5.1) and (5.2), let

$$R_i = \mathrm{sign}(\xi_i) * [I_{\xi_i = -1} Y_i^{-} + I_{\xi_i = 1} Y_i^{+}], \quad -\infty < i < \infty, \tag{5.3}$$

where the process $\{\xi_i\}$ is a simple Markov process which is independent of $\{Y_i^{\pm}\}$ and takes values in a finite set $\{-1, 0, 1\}$. $\{R_i\}$ is the return process.

Remark 6. The processes $\{\xi_i^{-}\}$, $\{\xi_i^{+}\}$ may be Bernoulli processes or Markov processes taking values in a finite set. The process $\{\xi_i\}$ may be considered as an independent process or a Markov process taking values in a finite set.

Remark 7. In Model (5.3), as long as $\{Y_i^{-}\}$, $\{Y_i^{+}\}$, and ξ_i are determined, R_i is determined.

Remark 8. In many applications, only positive observed values are concerned. Insurance claims, annual maxima of precipitations, file sizes, durations in internet traffic at certain point are some of those examples having positive values only. Even in our negative return model, the values have been converted into positive values. Therefore, the properties of Model (5.1) can easily be extended to Model (5.2). We now focus on Model (5.1).

We first compute the lag-k tail dependence index in Model (5.1):

$$\begin{aligned}\frac{P(R_j^- > u, R_{j+k}^- > u)}{P(R_j^- > u)} &= \frac{P(Y_j^- > u, \xi_j^- = 1, Y_{j+k}^- > u, \xi_{j+k}^- = 1)}{P(Y_j^- > u, \xi_j^- = 1)} \\ &= \frac{P(Y_j^- > u, Y_{j+k}^- > u)}{P(Y_j^- > u)} \frac{P(\xi_j^- = 1, \xi_{j+k}^- = 1)}{P(\xi_j^- = 1)} \\ &= \frac{P(Y_j^- > u, Y_{j+k}^- > u)}{P(Y_j^- > u)} P(\xi_{j+k}^- = 1 | \xi_j^- = 1) \\ &\to \min(a_{k0}^-, a_{kk}^-) P(\xi_{j+k}^- = 1 | \xi_j^- = 1) \end{aligned} \tag{5.4}$$

as u tends to infinite.

Now suppose $\{\xi_i^-\}$ is a simple Markov process taking values in a finite set $\{0, 1\}$, and with the transition probabilities:

$$\begin{aligned} P(\xi_{j+1}^- = 0 | \xi_j^- = 0) = p_{00}, P(\xi_{j+1}^- = 1 | \xi_j^- = 0) = p_{01}, \\ P(\xi_{j+1}^- = 0 | \xi_j^- = 1) = p_{10}, P(\xi_{j+1}^- = 1 | \xi_j^- = 1) = p_{11}, \end{aligned} \tag{5.5}$$

and the kth $(k > 1)$ step transition probabilities:

$$\begin{aligned} P(\xi_{j+k}^- = 0 | \xi_j^- = 0) = p_{00}^{(k)}, P(\xi_{j+k}^- = 1 | \xi_j^- = 0) = p_{01}^{(k)}, \\ P(\xi_{j+k}^- = 0 | \xi_j^- = 1) = p_{10}^{(k)}, P(\xi_{j+k}^- = 1 | \xi_j^- = 1) = p_{11}^{(k)}, \end{aligned} \tag{5.6}$$

where the superscripts (k) denote the kth step in a Markov process. Using (5.5) and (5.6), we first have

$$\begin{aligned} P(R_1^- < x, R_{1+r}^- < y) &= P(R_1^- < x, R_{1+r}^- < y, \xi_1^- = 0) \\ &\quad + P(R_1^- < x, R_{1+r}^- < y, \xi_1^- = 1) \\ &= P(R_{1+r}^- < y, \xi_1^- = 0) \\ &\quad + P(Y_1^- < x, R_{1+r}^- < y, \xi_1^- = 1). \end{aligned} \tag{5.7}$$

Since

$$\begin{aligned}P(R^-_{1+r}<y,\xi^-_1=0) &= P(R^-_{1+r}<y,\xi^-_{1+r}=0,\xi^-_1=0)+P(R^-_{1+r}<y,\xi^-_{1+r}=1,\xi^-_1=0)\\ &= P(\xi^-_{1+r}=0,\xi^-_1=0)+P(Y^-_{1+r}<y,\xi^-_{1+r}=1,\xi^-_1=0)\\ &= P(\xi^-_1=0)P(\xi^-_{1+r}=0|\xi^-_1=0)\\ &\quad + P(Y^-_{1+r}<y)P(\xi^-_1=0)P(\xi^-_{1+r}=1,\xi^-_1=0)\\ &= p_0p^{(r)}_{00}+p_0p^{(r)}_{01}e^{-1/y}, \end{aligned} \tag{5.8}$$

where $p_0 = P(\xi^-_1 = 0)$ is the probability of the chain starting at the initial state $\xi^-_1 = 0$;

$$\begin{aligned}P(Y^-_1<x,R^-_{1+r}<y,\xi^-_1=1) &= P(Y^-_1<x,R^-_{1+r}<y,\xi^-_{1+r}=0,\xi^-_1=1)\\ &\quad + P(Y^-_1<x,R^-_{1+r}<y,\xi^-_{1+r}=1,\xi^-_1=1)\\ &= P(Y^-_1<x,\xi^-_{1+r}=0,\xi^-_1=1)\\ &\quad + P(Y^-_1<x,Y^-_{1+r}<y)P(\xi^-_{1+r}=1,\xi^-_1=1)\\ &= p_1p^{(r)}_{10}e^{-1/x}+p_1p^{(r)}_{11}\exp\{-\frac{1}{x}-\frac{1}{y}\\ &\quad + \min(\frac{a^-_{r0}}{x},\frac{a^-_{rr}}{y})\}, \end{aligned} \tag{5.9}$$

where $p_1 = 1-p_0$; then putting (5.8) and (5.9) in (5.7), we have

$$\begin{aligned}P(R^-_1<x,R^-_{1+r}<y) &= p_0p^{(r)}_{00}+p_0p^{(r)}_{01}e^{-1/y}+p_1p^{(r)}_{10}e^{-1/x}\\ &\quad + p_1p^{(r)}_{11}\exp\{-\frac{1}{x}-\frac{1}{y}+\min(\frac{a^-_{r0}}{x},\frac{a^-_{rr}}{y})\}. \end{aligned} \tag{5.10}$$

Notice that the event $\{R^-_i<x\}$ is a union of two events $\{Y^-_i<x,\xi^-_i=1\}$ and $\{\xi^-_i=0\}$, which are mutually exclusive.Then for $I_1<I_2<\ldots<I_m$, we have the following joint probability expression:

$$\begin{aligned}&P(R^-_{I_1}<x_{I_1},R^-_{I_2}<x_{I_2},\ldots,R^-_{I_m}<x_{I_m})\\ &= p_1\prod_{j=1}^{m1}p^{(I_{j+1}-I_j)}_{11}P(Y^-_{I_1}<x_{I_1},Y^-_{I_2}<x_{I_2},\ldots,Y^-_{I_m}<x_{I_m})\\ &\quad + \sum_{k=1}^{m1}\Bigg[\sum_{\substack{j_1<j_2<\ldots<j_k\\ \{j_1j_2,\ldots j_k\}\subset\{I_1,I_2,\ldots,I_m\}}} p_0^{I_{(j_1>I_1)}}p_1^{I_{(j_1=I_1)}}\prod_{j=1}^{m1}p^{(I_{j+1}-I_j)}_{I_{(I_j\in\{j_1j_2,\ldots j_k\})}I_{(j+1\in\{j_1j_2,\ldots j_k\})}}\\ &\quad P(Y^-_{j_1}<x_{j_1},Y^-_{j_2}<x_{j_2},\ldots,Y^-_{j_k}<x_{j_k})\Bigg] + p_0\prod_{j=1}^{m1}p^{(I_{j+1}-I_j)}_{00}. \end{aligned} \tag{5.11}$$

Notice that an exact M4 data generating process (DGP) may not be observable in real – i.e., it may not be realistic to observe an infinite number of times of signature patterns as demonstrated in Fig. 5. It is natural to consider the following model:

$$R_i^* = \xi_i^-(Y_i^- + N_i^-), \quad -\infty < i < \infty, \tag{5.12}$$

where $\{N_i^-\}$ is an independent bounded noise process which is also independent of $\{\xi_i^-\}$ and $\{Y_i^-\}$. By adding the noise process, the signature patterns can not be explicitly illustrated as we did in Fig. 5.

The following proposition tells that as long as the characterization of tail dependencies is the main concern, Model (5.1) or its simplified form should be a good approximation to the possible true model.

Proposition 5.1. The lag-k tail dependence within $\{R_i^*\}$ can be expressed as:

$$\frac{P(R_j^* > u, R_{j+k}^* > u)}{P(R_j^* > u)} \to \min(a_{k0}^-, a_{kk}^-) P(\xi_{j+k}^- = 1 | \xi_j^- = 1)$$
$$as\ u \to \infty. \tag{5.13}$$

A proof of Proposition 5.1 is given in Section 9.

The same lag-k tail dependence index in both (5.4) and (5.13) suggests that an MCM3 process (5.1) can be used to approximate (5.12). Under (5.12), it is reasonable to assume each paired parameters being identical. Under this assumption, we derive the parameter estimators in the next Section.

6. MCM3 PARAMETER ESTIMATION

Notice that on the right-hand sides of (5.4) and (5.13) are moving coefficients and the rth step transition probabilities. We choose to substitute the quantities on the left-hand sides by their corresponding empirical counterparts to construct the parameter estimators.

Considering that $\{\xi_i^-\}$ and $\{Y_i^-\}$ are assumed independent, the estimations of parameters from these two processes can be processed separately. Our main interest here is to estimate the parameters in M3 process. The maximum likelihood estimations and the asymptotic normality of the estimators for Markov processes have been developed in Billingsley (1961a, b). Here we simply perform empirical estimation and then treat the estimated transition probabilities as known parameter values to accomplish statistical inference for the M3 processes.

We now assume $a_{r0} = a_{rr}$ for all $r = 1,\ldots,\ L = L^-$, and denote $P(R_i^- > u, R_{i+r}^- > u)$ as μ_r^- for $r = 1,...,L$, $P(R_1^- > u)$ as μ_{L+1}^-, a_{r0} as a_r respectively. Define

$$\bar{X}_r^- = \frac{1}{n}\sum_{i=1}^{n-r} I_{(R_i^- > u, R_{i+r}^- > u)}, \quad r = 1,\ldots,L, \tag{6.1}$$

$$\bar{X}_{L+1}^- = \frac{1}{n}\sum_{i=1}^{n-r} I_{(R_i^- > u)}. \tag{6.2}$$

Then by the strong law of large numbers (SLLN), we have

$$\bar{X}_r^- \xrightarrow{a.s.} P(R_i^- > u, R_{i+r}^- > u) = \mu_r^-, \quad r = 1,\ldots,L, \tag{6.3}$$

$$\bar{X}_{L+1} \xrightarrow{a.s.} P(R_1^- > u) = \mu_{L+1}^-. \tag{6.4}$$

From (5.4) and (5.13), we propose the following estimators for parameters a_r^-:

$$\hat{a}_r^- = \frac{\bar{X}_r^-}{\bar{X}_{L+1}^- p_{11}^{(r)}}, \quad r = 1,\ldots,L. \tag{6.5}$$

In order to study asymptotic normality, we introduce the following proposition which is Theorem 27.4 in Billingsley (1995). First we introduce the so-called α-mixing condition. For a sequence $Y_1,\ Y_2,\ldots$ of random variables, let α_n be a number such that

$$\left|P(A \cap B) - P(A)P(B)\right| \le \alpha_n$$

for $A \in \sigma(Y_1,\ldots,Y_k), B \in \sigma(Y_{k+n}, Y_{k+n+1},\ldots)$, and $k \ge 1, n \ge 1$. When $\alpha_n \to 0$, the sequence $\{Y_n\}$ is said to be α-mixing.

Proposition 6.1. Suppose that $X_1, X_2, \ldots$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $E[X_n] = 0$ and $E[X_n^{12}] < \infty$. If $S_n = X_1 + \cdots + X_n$, then

$$n^{-1}Var[S_n] \to \sigma^2 = E[X_1^2] + 2\sum_{k=1}^{\infty} E[X_1 X_{1+k}],$$

where the series converges absolutely. If $\sigma > 0$, then $S_n/\sigma\sqrt{n} \xrightarrow{\mathscr{L}} N(0,1)$.

Remark 9. The conditions $\alpha_n = O(n^{-5})$ and $E[X_n^{12}] < \infty$ are stronger than necessary as stated in the remark following Theorem 27.4 in Billingsley (1995) to avoid technical complication in the proof.

With the established notations, we have the following lemma dealing with the asymptotic properties of the empirical functions.

Lemma 6.2. Suppose that $\bar{X}_j^-$ and μ_j^- are defined in (6.1)–(6.4), then

$$\sqrt{n}\left(\begin{bmatrix}\bar{X}_1^- \\ \vdots \\ \bar{X}_{L+1}^-\end{bmatrix} - \begin{bmatrix}\mu_1^- \\ \vdots \\ \mu_{L+1}^-\end{bmatrix}\right) \xrightarrow{\mathscr{D}} N\left(0, \Sigma^- + \sum_{k=1}^{L}\{W_k^- + W_k^{-T}\}\right),$$

where the entries σ_{ij} of matrix Σ^- and the entries w_k^{ij} of the matrix W_k^- are defined below. For $i = 1, \ldots, L, j = 1, \ldots, L$,

$$\begin{aligned}\sigma_{ij} &= E(I_{\{R_1^- > u, R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_1^- > u, R_{1+j}^- > u\}} - \mu_j^-) \\ &= P(R_1^- > u, R_{1+i}^- > u, R_{1+j}^- > u) - \mu_i^- \mu_j^-.\end{aligned}$$

$$\begin{aligned}w_k^{ij} &= E(I_{\{R_1^- > u, R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u, \quad R_{1+j+k}^- > u\}} - \mu_j^-) \\ &= P(R_1^- > u, R_{1+i}^- > u, R_{1+k}^- > u, R_{1+j+k}^- > u) - \mu_i^- \mu_j^-.\end{aligned}$$

For $i = 1, \ldots, L, j = L+1$,

$$\begin{aligned}\sigma_{ij} &= E(I_{\{R_1^- > u, \quad R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_1^- > u\}} - \mu_j^-) \\ &= \mu_i^- - \mu_i^- \mu_j^-.\end{aligned}$$

$$\begin{aligned}w_k^{ij} &= E(I_{\{R_1^- > u, \quad R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u\}} - \mu_j^-) \\ &= P(R_1^- > u, \quad R_{1+i}^- > u, R_{1+k}^- > u) - \mu_i^- \mu_j^-.\end{aligned}$$

For $i = L+1, j = 1, \ldots, L$,

$$\sigma_{ij} = \sigma_{ji}.$$

$$\begin{aligned}w_k^{ij} &= E(I_{\{R_1^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u, R_{1+j+k}^- > u\}} - \mu_j^-) \\ &= P(R_1^- > u, R_{1+k}^- > u, R_{1+j+k}^- > u) - \mu_i^- \mu_j^-.\end{aligned}$$

For $i = L+1$, $j = L+1$,

$$\sigma_{ij} = \mu_i^- - \mu_i^- \mu_j^-.$$

$$\begin{aligned} w_k^{ij} &= E(I_{\{R_1^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u\}} - \mu_j^-) \\ &= P(R_1^- > u, R_{1+k}^- > u) - \mu_i^- \mu_j^- = \mu_k^- - \mu_i^- \mu_j^-. \end{aligned}$$

A proof of Lemma 6.2. is deferred to Section 9.

From (6.5), we have the following Jacobian matrix:

$$\Theta^- = \begin{pmatrix} \frac{1}{\mu_{L+1}^- p_{11}} & \cdots & \cdots & \cdots & \cdots & -\frac{\mu_1^-}{(\mu_{L+1}^-)^2 p_{11}} \\ & \frac{1}{\mu_{L+1}^- p_{11}^{(2)}} & & & & -\frac{\mu_2^-}{(\mu_{L+1}^-)^2 p_{11}^2} \\ & & & \ddots & & \\ & & & & \frac{1}{\mu_{L+1}^- p_{11}^{(L)}} & -\frac{\mu_L^-}{(\mu_{L+1}^-)^2 p_{11}^{(L)}} \end{pmatrix} \tag{6.6}$$

which is applied to getting asymptotic covariance matrix for parameter estimators.

We have obtained the following theorem.

Theorem 6.3. Let $\hat{a}_0^- = 1 - 2\sum_{i=1}^L \hat{a}_i^-$ Let $\hat{\mathbf{a}}^- = [\hat{a}_0^-, \hat{a}_1^-, \cdots, \hat{a}_L^-]^T, \mathbf{a}^- = [a_0^-, a_1^-, \cdots, a_L^-]^T$ be two vectors. Then with the established notations,

$$\sqrt{n}(\hat{\mathbf{a}}^- - \hat{\mathbf{a}}^-) \xrightarrow{\mathscr{L}} N(0, C^- B^- C^{-T}),$$

where B^- is a matrix with elements $B_{1j}^- = 0$, $B_{i1}^- = 0$, $I, j = 1,2,\ldots, m$, and the minor of B_{11}^- being $\Theta^-[\Sigma^- + \sum_{k=1}^L \{W_k^- + W_k^{-T}\}]\Theta^{-T}$; and C^- is a matrix with elements $C_{11}^- = 1$, $C_{1j}^- = -2$, $C_{i1}^- = 0$, $I, j = 2,\ldots,$m, and the minor of C_{11}^- being a unit matrix.

Similarly, we can construct estimators of parameters in Model 2. For Model 3, we propose to apply Model 1 and Model 2 first, and then re-estimate the one step Markov transition probabilities. The asymptotic properties of the transition probability estimators can be found in Billingsley (1961a, b). We also can derive the joint asymptotic covariance matrix for all parameter estimators following the similar process used for the combined M3 and Markov process for negative returns. We will not pursue that in this paper because we think the joint asymptotic properties from Model 1 and Model 2 may be enough for various practical purposes.

7. MODELING JUMPS IN RETURNS

7.1. Modeling Jumps in Negative Returns

The previous analysis in Section 3 has suggested up to lag-12 tail dependencies for the negative. We now fit model (5.1) to the transformed negative returns.

From the data, the proportion of the days that the negative returns are observed, i.e. the investors lose money, is 0.4731. We use Markov chain to model negative signs. Suppose the state 1 corresponds to the day that a negative return is observed and the state 0 corresponds to the day that a negative return is not observed. The one step transition probabilities $P(\xi_d = I|\xi_d = j)$, $I, j = 0,1$, are estimated in the following table.

State	0	1
0	0.5669	0.4331
1	0.4823	0.5177

They are empirical estimations – for example, $P(\xi_{i+1,d} = 0|\xi_{id} = 0) = P(Y_{i+1,d} \leq 0|Y_{i,d} \leq 0)$ is estimated by: $\sum_{i=1}^{n-1} I_{(Y_{i,d}\leq 0, Y_{i+1,d}\leq 0)} \Big/ \sum_{i=1}^{n-1} I_{(Y_{i,d}\leq 0)}$. The following table estimates the rth step transition probabilities using the data and computes the transition probabilities using Chapman-Kolmogorov equation.

Step	Estimated				Chapman–Kolmogorov			
1	0.5669	0.4331	0.4823	0.5177	0.5669	0.4331	0.4823	0.5177
2	0.5172	0.4828	0.5377	0.4623	0.5303	0.4697	0.5231	0.4769
3	0.5195	0.4805	0.5353	0.4647	0.5272	0.4728	0.5266	0.4734
4	0.5302	0.4698	0.5232	0.4768	0.5269	0.4731	0.5269	0.4731
5	0.5284	0.4716	0.5251	0.4749	0.5269	0.4731	0.5269	0.4731
6	0.5223	0.4777	0.5319	0.4681	0.5269	0.4731	0.5269	0.4731
7	0.5279	0.4721	0.5256	0.4744	0.5269	0.4731	0.5269	0.4731
8	0.5282	0.4718	0.5256	0.4744	0.5269	0.4731	0.5269	0.4731
9	0.5239	0.4761	0.5301	0.4699	0.5269	0.4731	0.5269	0.4731
10	0.5285	0.4715	0.5252	0.4748	0.5269	0.4731	0.5269	0.4731
11	0.5270	0.4730	0.5271	0.4729	0.5269	0.4731	0.5269	0.4731
12	0.5369	0.4631	0.5158	0.4842	0.5269	0.4731	0.5269	0.4731

One can see from the table that the estimated values (the middle panel) are very close to the theoretical values (the right panel) after the first step. The limiting distribution of the two state Markov chain is consistent with the proportions of the days that a negative return is observed. Based on this table, one can conclude that the data suggests that a two state Markov chain model is a good fit of the transitions of signs of negative returns. This analysis together with the previous analysis suggest that an MCM3 model is suitable for jumps in returns. Next we estimate the parameters in M3 model. The results are summarized in Table 3. The estimated parameter values can be used to further statistical inference – for example, to compute value at risk (VaR) based on the estimated model.

7.2. Modeling Jumps in Positive Returns

Similar to the case of negative returns, the estimated first order transition probability matrix for positive returns is summarized in the following table.

State	0	1
0	0.5209	0.4791
1	0.4379	0.5621

The following table estimates the rth step transition probabilities using the data and computes the transition probabilities using Chapman–Kolmogorov equation.

Step	Estimated				Chapman-Kolmogorov			
1	0.5209	0.4791	0.4379	0.5621	0.5209	0.4791	0.4379	0.5621
2	0.4649	0.5351	0.4893	0.5107	0.4811	0.5189	0.4742	0.5258
3	0.4665	0.5335	0.4876	0.5124	0.4778	0.5222	0.4773	0.5227
4	0.4804	0.5196	0.4749	0.5251	0.4776	0.5224	0.4775	0.5225
5	0.4786	0.5214	0.4766	0.5234	0.4775	0.5225	0.4775	0.5225
6	0.4722	0.5278	0.4825	0.5175	0.4775	0.5225	0.4775	0.5225
7	0.4782	0.5218	0.4772	0.5228	0.4775	0.5225	0.4775	0.5225
8	0.4791	0.5209	0.4762	0.5238	0.4775	0.5225	0.4775	0.5225

The estimated parameter values in M3 model are summarized in Table 4. The estimated parameter values can be used to further statistical inference – for

Table 3. Estimations of Parameters in Model (5.1).

r	a_{r0}	SE	r	a_{r0}	SE	r	a_{r0}	SE
0	0.2385	0.2318						
1	0.0700	0.0175	5	0.0306	0.0178	9	0.0175	0.0176
2	0.0700	0.0181	6	0.0088	0.0176	10	0.0131	0.0176
3	0.0525	0.0179	7	0.0131	0.0176	11	0.0088	0.0176
4	0.0525	0.0180	8	0.0306	0.0177	12	0.0131	0.0176

Table 4. Estimations of Parameters in Model (5.2).

r	a_{r0}	SE	r	a_{r0}	SE
0	0.5913	0.1668	4	0.0314	0.0200
1	0.0550	0.0190	5	0.0236	0.0199
2	0.0354	0.0200	6	0.0118	0.0198
3	0.0314	0.0200	7	0.0157	0.0199

example, to compute value at risk (VaR) based on the estimated model. However, we restrict ourself to find estimates of the M3 process in this current work.

8. DISCUSSION

In this paper, we obtained statistical evidences of the existence of extreme impacts in financial time series data; we introduced a new time series model structure – i.e., combinations of Markov processes, GARCH(1,1) volatility model, and M3 processes. We restricted our attentions to a subclass of M3 processes. This subclass has advantages of efficiently modeling serial tail dependent financial time series, while, of course, other model specifications are possibly also suitable.

We proposed models for tail dependencies in jumps in returns. The next step is to make statistical and economic inferences of computing risk measures and constructing prediction intervals. In a different project, we extend the results developed in this paper to study extremal risk analysis and portfolio choice.

The approach adopted in the paper is a hierarchical model structure – i.e., to apply GARCH(1,1) fitting and to get estimated standard deviations first; then based on standardized return series, we apply M3 and Markov

processes modelling. It is possible to study Markov processes, GARCH processes, and M3 processes simultaneously. But that requires additional work. We put this goal as a direction for future research.

NOTE

1. Matlab codes for the gamma test are available upon request.

ACKNOWLEDGEMENT

This work was supported in part by NSF Grants DMS-0443048 and DMS-0505528, and Institute for Mathematical Research, ETH Zürich. The author thanks Professor Paul Embrechts, Professor Tom Fomby (Editor), and one referee for their many comments, suggestions which have greatly improved the quality of the paper.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC–19, 716–723.

Billingsley, P. (1961a). *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.

Billingsley, P. (1961b). Statistical methods in Markov chains. *Annals of Mathematical Statistics*, *32*, 12–40.

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: Wiley.

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

Danielsson, J. (2002). The emperor has no clothes: Limits to risk modelling. *Journal of Banking and Finance*, *26*, 1273–1296.

Davis, R. A., & Resnick, S. I. (1989). Basic properties and prediction of max-ARMA processes. *Advances in Applied Probability*, *21*, 781–803.

Davis, R. A., & Resnick, S. I. (1993). Prediction of stationary max-stable processes. *Annals of Applied Probability*, *3*, 497–525.

Deheuvels, P. (1983). Point processes and multivariate extreme values. *Journal of Multivariate Analysis*, *13*, 257–272.

Duan, J. C., Ritchken, P., & Sun, Z. (2003). Option valuation with jumps in returns and volatility. Web reference.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Berlin: Springer.

Embrechts, P., McNeil, A., & Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In: M. A. H. Dempster (Ed.), *Risk management: Value at risk and beyond* (pp. 176–223). Cambridge: Cambridge University Press.

Engle, R. (2002). Dynamic conditional correlation – a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics*, *17*, 339–350.

Eraker, B., Johannes, M., & Polson, N. G. (2003). The impact of jumps in returns and volatility. *Journal of Finance*, *53*, 1269–1300.

de Haan, L. (1984). A spectral representation for max-stable processes. *Annals of Probability*, *12*, 1194–1204.

de Haan, L., & Resnick, S. I. (1977). Limit theory for multivariate sample extremes. *Zeitschriftfur Wahrscheinlichkeitstheorieund verwandle. Gebiete*, *40*, 317–337.

Hall, P., Peng, L., & Yao, Q. (2002). Moving-maximum models for extrema of time series. *Journal of Statistical Planning and Inference*, *103*, 51–63.

Hall, P., & Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica*, *71*, 285–317.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Berlin: Springer.

McNeil, A., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, *7*, 271–300.

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

Mikosch, T., & Straumann, D. (2002). Whittle estimation in a heavy-tailed GARCH(1,1) model. *Stochastic Process. Appl*, *100*, 187–222.

Nandagopalan, S. (1990). *Multivariate extremes and the estimation of the extremal index*. Ph.D. dissertation, Dept. of Statistics, University of North Carolina, Chapel Hill.

Nandagopalan, S. (1994). On the multivariate extremal index. *Journal of Research, National Institute of Standards and Technology*, *99*, 543–550.

Pickands, J., III. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, *3*, 119–131.

Resnick, S. I. (1987). *Extreme values, regular variation, and point processes*. New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sibuya, M. (1960). Bivariate extreme statistics, I. *Annals of Institute of Statistical Mathematics*, *1*, 195–210.

Smith, R. L. (2003). Statistics of extremes: With applications in the environment, insurance and finance. In: B. Finkenstadt & H. Rootzén (Eds), *Extreme value in finance, Telecommunications and the Environment* (pp. 1–78). Chapman & Hall/CRC: Boca Raton.

Smith, R. L., & Weissman, I. (1996). *Characterization and estimation of the multivariate extremal index*. Technical report, the University of North Carolina, USA.

Straumann, D. (2003). *Estimation in conditionally heteroscedastic time series models*. Ph.D. thesis, Laboratory of Actuary Mathematics, University of Copenhagen.

Tsay, R. S. (1999). *Extreme value analysis of financial data*. Manuscript, Univ. of Chicago.

Zhang, Z. (2003a). *Quotient correlation*: *A sample based alternative to Pearson's correlation*. Manuscript, Washington University.

Zhang, Z. (2003b). *The estimation of M4 processes with geometric moving patterns*. Manuscript, Washington University.

Zhang, Z. (2004). *Some results on approximating max-stable processes*. Manuscript, Washington University.

Zhang, Z. & Smith, R. L. (2003). *On the estimation and application of max-stable processes*. Manuscript, Washington University.

Zhang, Z., & Smith, R. L. (2004). The behavior of multivariate maxima of moving maxima processes. *Journal of Applied Probability*, *41*, 1113–1123.

APPENDIX

Proof of Proposition 5.1. We have

$$\begin{aligned} P(R_j^* > u, R_{j+k}^* > u) &= P(Y_j^- + N_j^- > u, Y_{j+k}^- + N_{j+k}^- > u, \xi_j^- = 1, \xi_{j+k}^- = 1) \\ &= p_1 p_{11}^{(k)} P(Y_j^- + N_j^- > u, Y_{j+k}^- + N_{j+k}^- > u), \\ P(R_j^* > u) &= p_1 P(Y_j^- + N_j^- > u)\cdot \end{aligned}$$

Let $f(x)$ and M be the density and the bound limit of N_j, then

$$\begin{aligned} \frac{P(R_j^* > u, R_{j+k}^* > u)}{P(R_j^* > u)} &= p_{11}^{(k)} \frac{P(Y_j^- + N_j^- > u, Y_{j+k}^- + N_{j+k}^- > u)}{P(Y_j^- + N_j^- > u)} \\ &= p_{11}^{(k)} \frac{\int_{-M}^{M}\int_{-M}^{M} P(Y_j^- > u - x, Y_{j+k}^- > u - y) f(x) f(y) \mathrm{d}x \mathrm{d}y}{\int_{-M}^{M} p(Y_j^- > u - x) f(x) \mathrm{d}x} \\ &= p_{11}^{(k)} \frac{\int_{-M}^{M}\int_{-M}^{M} \frac{P(Y_j^- > u-x, Y_{j+k}^- > u-y)}{P(Y_j^- > u)} f(x) f(y) \mathrm{d}x \mathrm{d}y}{\int_{-M}^{M} \frac{P(Y_j^- > u-x)}{P(Y_j^- > u)} f(x) \mathrm{d}x}. \end{aligned}$$

It is easy to see that $\lim_{u\to\infty} \frac{P(Y_j^- > u-x)}{P(Y_j^- > u)} = 1$, and

$$\frac{P(Y_j^- > u - x, Y_{j+k}^- > u - y)}{P(Y_j^- > u)} = \frac{P(Y_j^- > u - x)}{P(Y_j^- > u)} \frac{P(Y_j^- > u - x, Y_{j+k}^- > u - y)}{P(Y_j^- > u - x)}.$$

We have

$$\begin{aligned} \frac{P(Y_j^- > w, Y_{j+k}^- > w + z)}{P(Y_j^- > w)} &= \frac{1 - e^{-1/w} - e^{-1/(w+z)} + e^{-1/w - 1/(w+z) + \min(a_{k0}/w, a_{kk}/(w+z))}}{1 - e^{-1/w}} \\ &= 1 - e^{-1/(w+z)} \left[\frac{1 - e^{-1/w + \min(a_{k0}/w, a_{kk}/(w+z))}}{1 - e^{-1/w}}\right]. \end{aligned}$$

Since for $w>0$, $z>0$ (similarly for $z<0$), we have

$$\min(a_{k0}/(w+z), a_{kk}/(w+z)) \le \min(a_{k0}/w, a_{kk}/(w+z)) \le \min(a_{k0}/w, a_{kk}/w),$$

so

$$\frac{1-e^{-1/w+\min(a_{k0},a_{kk})/w}}{1-e^{-1/w}} \le \frac{1-e^{-1/w+\min(a_{k0}/w,a_{kk}/(w+z))}}{1-e^{-1/w}} \le \frac{1-e^{-1/w+\min(a_{k0},a_{kk})/(w+z)}}{1-e^{-1/w}}$$

which gives

$$\lim_{w\to\infty} \frac{1-e^{-1/w+\min(a_{k0},a_{kk})/w}}{1-e^{-1/w}} = 1-\min(a_{k0}, a_{kk}),$$

$$\lim_{w\to\infty} \frac{1-e^{-1/w+\min(a_{k0},a_{kk})/(w+z)}}{1-e^{-1/w}} = 1-\min(a_{k0}, a_{kk}),$$

hence,

$$\lim_{w\to\infty} \frac{1-e^{-1/w+\min(a_{k0}/w,a_{kk}/(w+z))}}{1-e^{-1/w}} = 1-\min(a_{k0}, a_{kk}),$$

So we have

$$\lim_{w\to\infty} \frac{P(Y_j^- > w, Y_{j+k}^- > w+z)}{P(Y_j^- > w)} = \min(a_{k0}, a_{kk}).$$

Let $w = u - x$, $z = x - y$, then

$$\lim_{u\to\infty} \frac{P(Y_j^- > u-x, Y_{j+k}^- > u-y)}{P(Y_j^- > u)} = \min(a_{k0}, a_{kk})$$

which gives the desired results in (5.13). *Proof* of Lemma 6.2. Let

$$U_1 = (I_{(R_1^- > u, R_2^- > u)} - \mu_1^-, I_{(R_1^- > u, R_3^- > u)} - \mu_2^-, \ \dots \ , I_{(R_1^- > u, R_{1+L}^- > u)} - \mu_L^-, I_{(R_1^- > u)} - \mu_{1+L}^-)^T,$$

$$U_{1+k} = (I_{(R_{1+k}^- > u, R_{2+k}^- > u)} - \mu_1^-, I_{(R_{1+k}^- > u, R_{3+k}^- > u)} - \mu_2^-, \ \dots \ , I_{(R_{1+k}^- > u, R_{1+k+L}^- > u)} - \mu_L^-, I_{(R_{1+k}^- > u)} - \mu_{1+L}^-)^T,$$

and $\alpha = (\alpha_1, \dots, \alpha_L)^T \neq 0$ be an arbitrary vector.

Let $X_1 = \alpha^T U_1, X_{1+k} = \alpha^T U_{1+k}, \dots$, then $E[X_n] = 0$ and $E[X_n^{12}] < \infty$. So Proposition 6.1 can apply. We say expectation are applied on all elements if expectation is applied on a random matrix. But $E[X_1^2] = \alpha^T E[U_1 U_1^T]\alpha =$

$\alpha^T\Sigma\alpha$, $E[X_1X_{1+k}] = \alpha^T E[U_1U_{1+k}^T]\alpha = \alpha^T W_k\alpha$. The entries of Σ and W_k are computed from the following expressions.

For $i = 1, \ldots, L$, $j = 1, \ldots, L$,

$$\begin{aligned}\sigma_{ij} &= E(I_{\{R_1^- > u, R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_1^- > u, R_{1+j}^- > u\}} - \mu_j^-)\\ &= P(R_1^- > u, R_{1+i}^- > u, R_{1+j}^- > u) - \mu_i^-\mu_j^-,\\ w_k^{ij} &= E(I_{\{R_1^- > u, R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u,\quad R_{1+j+k}^- > u\}} - \mu_j^-)\\ &= P(R_1^- > u, R_{1+i}^- > u, R_{1+k}^- > u, R_{1+j+k}^- > u) - \mu_i^-\mu_j^-.\end{aligned}$$

For $i = 1, \ldots, L$, $j = L+1$,

$$\begin{aligned}\sigma_{ij} &= E(I_{\{R_1^- > u, R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_1^- > u\}} - \mu_j^-)\\ &= \mu_i^- - \mu_i^-\mu_j^-,\\ w_k^{ij} &= E(I_{\{R_1^- > u, R_{1+i}^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u\}} - \mu_j^-)\\ &= P(R_1^- > u, R_{1+i}^- > u, R_{1+k}^- > u) - \mu_i^-\mu_j^-.\end{aligned}$$

For $i = L+1$, $j = 1, \ldots, L$,

$$\sigma_{ij} = \sigma_{ji},$$

$$\begin{aligned}w_k^{ij} &= E(I_{\{R_1^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u, R_{1+j+k}^- > u\}} - \mu_j^-)\\ &= P(R_1^- > u, R_{1+k}^- > u, R_{1+j+k}^- > u) - \mu_i^-\mu_j^-.\end{aligned}$$

For $i = L+1$, $j = L+1$,

$$\sigma_{ij} = \mu_i^- - \mu_i^-\mu_j^-,$$

$$\begin{aligned}w_k^{ij} &= E(I_{\{R_1^- > u\}} - \mu_i^-)(I_{\{R_{1+k}^- > u\}} - \mu_j^-)\\ &= P(R_1^- > u, R_{1+k}^- > u) - \mu_i^-\mu_j^- = \mu_k^- - \mu_i^-\mu_j^-.\end{aligned}$$

So the proof is completed by applying the Cramér–Wold device.