Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES

Time Series and Related Topics

In Memory of Ching-Zong Wei

Hwai-Chung Ho, Ching-Kang Ing, Tze Leung Lai, Editors

Volume 52

ISBN 978-0-940600-68-3
ISBN 0-940600-68-4
ISSN 0749-2170

Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES

Volume 52

Time Series and Related Topics

In Memory of Ching-Zong Wei

Hwai-Chung Ho, Ching-Kang Ing, Tze Leung Lai, Editors

Institute of Mathematical Statistics
Beachwood, Ohio, USA

Institute of Mathematical Statistics
Lecture Notes–Monograph Series

Series Editor:
R. A. Vitale

The production of the *Institute of Mathematical Statistics Lecture Notes–Monograph Series* is managed by the IMS Office: Jiayang Sun, Treasurer and Elyse Gustafson, Executive Director.

Library of Congress Control Number: 2006936508

International Standard Book Number (13): 978-0-940600-68-3

International Standard Book Number (10): 0-940600-68-4

International Standard Serial Number: 0749-2170

Printed in the United States of America

Contents

Contributors to this volume v

Preface
Hwai-Chung Ho, Ching-Kang Ing and Tze Leung Lai vii

CHING-ZONG WEI: BIOGRAPHICAL SKETCH AND BIBLIOGRAPHY

Biographical sketch ix

Bibliography x

Photographs xiv

ESTIMATION AND PREDICTION IN TIME SERIES MODELS

Pile-up probabilities for the Laplace likelihood estimator of a non-invertible first order moving average
F. Jay Breidt, Richard A. Davis, Nan-Jung Hsu and Murray Rosenblatt 1

Recursive estimation of possibly misspecified MA(1) models: Convergence of a general algorithm
James L. Cantor and David F. Findley 20

Estimation of AR and ARMA models by stochastic complexity
Ciprian Doru Giurcăneanu and Jorma Rissanen 48

On prediction errors in regression models with nonstationary regressors
Ching-Kang Ing and Chor-Yiu Sin 60

Forecasting unstable processes
Jin-Lung Lin and Ching-Zong Wei 72

Order determination in general vector autoregressions
Bent Nielsen 93

The distribution of model averaging estimators and an impossibility result regarding its estimation
Benedikt M. Pötscher 113

Conditional-sum-of-squares estimation of models for stationary time series with long memory
P. M. Robinson 130

TIME SERIES MODELING IN FINANCE, MACROECONOMICS AND OTHER APPLICATIONS

Modeling macroeconomic time series via heavy tailed distributions
J. A. D. Aston 138

Fractional constant elasticity of variance model
Ngai Hang Chan and Chi Tim Ng 149

Estimation errors of the Sharpe ratio for long-memory stochastic volatility models
Hwai-Chung Ho 165

Cowles commission structural equation approach in light of nonstationary time series analysis
Cheng Hsiao 173

Combining domain knowledge and statistical models in time series analysis
Tze Leung Lai and Samuel Po-Shing Wong 193

Multivariate volatility models
Ruey S. Tsay 210

RELATED TOPICS

Multi-armed bandit problem with precedence relations
Hock Peng Chan, Cheng-Der Fuh and Inchi Hu 223

Poisson process approximation: From Palm theory to Stein's method
Louis H. Y. Chen and Aihua Xia 236

Statistical modeling for experiments with sliding levels
Shao-Wei Cheng, C. F. J. Wu and Longcheen Huwang 245

Price systems for markets with transaction costs and control problems for some finance problems
Tzuu-Shuh Chiang, Shang-Yuan Shiu and Shuenn-Jyi Sheu 257

A note on the estimation of extreme value distributions using maximum product of spacings
T. S. T. Wong and W. K. Li 272

Some results on the Gittins index for a normal reward process
Yi-Ching Yao 284

Contributors to this volume

Aston, J. A. D. *Academia Sinica*

Breidt, F. J. *Colorado State University*

Cantor, J. L. *Science Application International Corporation*
Chan, H. P. *National University of Singapore*
Chan, N. H. *The Chinese University of Hong Kong*
Chen, L. H. Y. *National University of Singapore*
Chiang, T.-S. *Academia Sinica*

Davis, R. A. *Colorado State University*

Findley, D. F. *U.S. Census Bureau*
Fuh, C.-D. *Academia Sinica*

Giurcăneanu, C. D. *Tampere University of Technology*

Ho, H.-C. *Academia Sinica*
Hsiao, C. *University of Southern California*
Hsu, N.-J. *National Tsing-Hua University*
Hu, I. *Hong Kong University of Science and Technology*

Ing, C.-K. *Academia Sinica*

Lai, T. L. *Stanford University*
Li, W. K. *The University of Hong Kong*
Lin, J.-L. *Academia Sinica*

Ng, C. T. *The Chinese University of Hong Kong*
Nielsen, B. *University of Oxford*

Pötscher, B. M. *University of Vienna*

Rissanen, J. *Technical University of Tampere and Helsinki, and Helsinki Institute for Information Technology*
Robinson, P. M. *London School of Economics*
Rosenblatt, M. *University of California at San Diego*

Sheu, S.-J. *Academia Sinica*
Shiu, S.-Y. *University of Utah*
Sin, C.-Y. *Xiamen University*

Tsay, R. S. *University of Chicago*

Wei, C.-Z. *Academia Sinica*
Wong, S. P.-S. *The Chinese Universty of Hong Kong*
Wong, T. S. T. *The University of Hong Kong*

Xia, A. *University of Melbourne*

Yao, Y.-C. *Academia Sinica*

Preface

A major research area of Ching-Zong Wei (1949–2004) was time series models and their applications in econometrics and engineering, to which he made many important contributions. A conference on time series and related topics in memory of him was held on December 12–14, 2005, at Academia Sinica in Taipei, where he was Director of the Institute of Statistical Science from 1993 to 1999. Of the forty-two speakers at the conference, twenty contributed to this volume. These papers are listed under the following three headings.

1. Estimation and prediction in time series models

Breidt, Davis, Hsu and Rosenblatt consider estimation of the unknown moving average parameter θ in an MA(1) model when $\theta = 1$, and derive the limiting pile-up probabilities $P(\hat{\theta} = 1)$ and $1/n$-asymptotics for the Laplace likelihood estimator $\hat{\theta}$. Cantor and Findley introduce a recursive estimator for θ in a possibly misspecified MA(1) model and obtain convergence results by approximating the recursive algorithm for the estimator by a Robbins–Monro-type stochastic approximation scheme. Giurcăneanu and Rissanen consider estimation of the order of AR and ARMA models by stochastic complexity, which is the negative logarithm of a normalized maximum likelihood universal density function. Nielsen investigates estimation of the order in general vector autoregressive models and shows that likelihood-based information criteria, and likelihood ratio tests and residual-based tests can be used, regardless of whether the characteristic roots are inside, or on, or outside the unit disk, and also in the presence of deterministic terms. Instead of model selection, Pötscher considers model averaging in linear regression models, and derives the finite-sample and asymptotic distributions of model averaging estimators. Robinson derives the asymptotic properties of conditional-sum-of squares estimates in parametric models of stationary time series with long memory. Ing and Sin consider the final prediction error and the accumulated prediction error of the adaptive least squares predictor in stochastic regression models with nonstationary regressors. The paper by Lin and Wei, which was in preparation when Ching-Zong was still healthy, investigates the adaptive least squares predictor in unit-root nonstationary processes.

2. Time series modeling in finance, macroeconomics and other applications

Aston considers criteria for deciding when and where heavy-tailed models should be used for macroeconomic time series, especially those in which outliers are present. Hsiao reviews nonstationary time series analysis from the perspective of the Cowles Commission structural equation approach, and shows that the same rank condition for identification holds for both stationary and nonstationary time series, that certain instrumental variables are needed for consistent parameter estimation, and that classical instrumental-variable estimators have to be modified for valid inference in the presence of unit roots. Chan and Ng investigate option pricing when

the volatility of the underlying asset follows a fractional version of the CEV (constant elasticity of variance) model. Ho considers linear process models, with a latent long-memory volatility component, for asset returns and provides asymptotically normal estimates, with a slower convergence rate than $1/\sqrt{n}$, of the Sharpe ratios in these investment models. Tsay reviews some commonly used models for the time-varying multivariate volatility of k (≥ 2) assets and proposes a simple parsimonious approach that satisfies positive definite constraints on the time-varying correlation matrix. Lai and Wong propose a new approach to time series modeling that combines subject-matter knowledge of the system dynamics with statistical techniques in time series analysis and regression, and apply this approach to American option pricing and the Canadian lynx data.

3. Related topics

Besides time series analysis, Ching-Zong also made important contributions to the multi-armed bandit problem, estimation in branching processes with immigration, stochastic approximation, adaptive control and limit theorems in probability, and had an active interest in the closely related areas of experimental design, stochastic control and estimation in non-regular and non-ergodic models. The paper by Chan, Fu and Hu uses the multi-armed bandit problem with precedence relations to analyze a multi-phase management problem and thereby establishes the asymptotic optimality of certain strategies. Yao develops an approximation to Gittins index in the discounted multi-armed bandit problem by using a continuity correction in an associated optional stopping problem. Chen and Xia describe Stein's method for Poisson approximation and for Poisson process approximation from the points of view of immigration-death processes and Palm distributions. Cheng, Wu and Huwang propose a new approach, which is based on a response surface model, to the analysis of experiments that use the technique of sliding levels to treat related factors, and demonstrate the superiority of this approach over previous methods in the literature. Chiang, Sheu and Shiu formulate the valuation problem of a financial derivative in markets with transaction costs as a stochastic control problem and consider optimization of expected utility by using the price systems for these markets. Wong and Li propose to use the maximum product of spacings (MPS) method for parameter estimation in the GEV (generalized extreme value) family and the generalized Pareto family of distributions, and show that the MPS estimates are asymptotically efficient and can outperform the maximum likelihood estimates.

We thank the Institute of Statistical Science of Academia Sinica for providing financial support for the conference. Special thanks also go to the referees who reviewed the manuscripts. A biographical sketch of Ching-Zong and a bibliography of his publications appear after this Preface.

Hwai-Chung Ho
Ching-Kang Ing
Tze Leung Lai

Biographical sketch

Ching-Zong Wei was born in 1949 in south Taiwan. He studied mathematics at National Tsing-Hua University, Taiwan, where he earned a BS degree in 1971 and an MS degree in 1973. He went to the United States in 1976 to pursue advanced studies in statistics at Columbia University, where he earned a PhD degree in 1980. He then joined the Department of Mathematics at the University of Maryland, College Park, as an Assistant Professor in 1980, and was promoted to Associate Professor in 1984 and Full Professor in 1988. In 1990 he returned to Taiwan, his beloved homeland, to join the Institute of Statistical Science at Academia Sinica, where he stayed as Research Fellow for the rest of his life, serving between 1993 and 1999 as Director of the Institute. He also held a joint appointment with the Department of Mathematics at National Taiwan University.

In addition to his research and administrative work at Academia Sinica, Ching-Zong also made important contributions to statistical education in Taiwan. To promote statistical thinking among the general public, he published in local newspapers and magazines articles on various topics of general interest such as lottery games and the Bible code. These articles, written in Chinese, introduced basic statistical and probabilistic concepts in a heuristic and reader-friendly manner via entertaining stories, without formal statistical jargon.

Ching-Zong made fundamental contributions to stochastic regression, adaptive control, nonstationary time series, model selection and sequential design. In particular, his pioneering works on (i) strong consistency of least squares estimates in stochastic regression models, (ii) asymptotic behavior of least squares estimates in unstable autoregressive models, and (iii) predictive least squares principles in model selection, have been influential in control engineering, econometrics and time series. A more detailed description of his work appears in the Bibliography. He was elected Fellow of the Institute of Mathematical Statistics in 1989, and served as an Associate Editor of the *Annals of Statistics* (1987–1993) and *Statistic Sinica* (1991–1999). In 1999, when Ching-Zong was at the prime of his career, he was diagnosed with brain tumors. He recovered well after the first surgery and remained active in research and education. In 2002, he underwent a second surgery after recurrence of the tumors, which caused deterioration of his vision. He continued his work and courageous fight with brain tumors and passed away on November 18, 2004, after an unsuccessful third surgery. He was survived by his wife of close to 30 years, Mei, and a daughter. In recognition of his path-breaking contributions, Vol. 16 of *Statistica Sinica* contains a special memorial section dedicated to him.

Bibliography

Before listing Ching-Zong's publications, we give a brief introduction of their background and divide them broadly into five groups, in which the papers are referred to by their numbers in the subsequent list.

A. Least squares estimates in stochastic regression models

Ching-Zong's work in this area began with papers [1], [2] and [3], in which the strong consistency of least squares estimates is established in fixed-design linear regression models. In particular, when the errors are square integrable martingale differences, a necessary and sufficient condition for the strong consistency of least squares estimates is given. However, when the regressors are stochastic, this condition is too weak to ensure consistency. Paper [6] is devoted to resolving this difficulty, and establishes strong consistency and asymptotic normality of least squares estimates in stochastic regression models under mild assumptions on the stochastic regressors and errors. These results can be applied to interval estimation of the regression parameters and to recursive on-line identification and control schemes for linear dynamic systems, as shown in [6]. Papers [7], [12] and [15] extend the results of [6] and establish the asymptotic properties of least squares estimates in more general settings.

B. Adaptive control and stochastic approximation

Papers [17] and [18] resolve the dilemma between the control objective and the need of information for parameter estimation by occasional use of white-noise probing inputs and by a reparametrization of the model. Asymptotically efficient self-tuning regulators are constructed in [18] by making use of certain basic properties of adaptive predictors involving recursive least squares for the reparametrized model. Paper [16] studies excitation properties of the designs generated by adaptive control schemes. Instead of using least squares, [13] uses stochastic approximation for recursive estimation of the unknown parameters in adaptive control. Paper [20] introduces a multivariate version of adaptive stochastic approximation and demonstrates that it is asymptotically efficient from both the estimation and control points of view, while [28] uses martingale transforms with non-atomic limits to analyze stochastic approximation. Paper [23] introduces irreversibility constraints into the classical multi-armed bandit problem in adaptive control.

C. Nonstationary time series

For a general autoregressive (AR) process, [9] proves for the first time that the least squares estimate is strongly consistent regardless of whether the roots of the characteristic polynomial lie inside, on, or outside the unit disk. Paper [22] shows that in general unstable AR models, the limiting distribution of the least squares estimate can be characterized as a function of stochastic integrals. The techniques

developed in [22] and in the earlier paper [19] for deriving the asymptotic distribution soon became standard tools for analyzing unstable time series and led to many important developments in econometric time series, including recent advances in the analysis of cointegration processes.

D. Adaptive prediction and model selection

Paper [21] considers sequential prediction problems in stochastic regression models with martingale difference errors, and gives an asymptotic expression for the cumulative sum of squared prediction errors under mild conditions. Paper [27] shows that Rissanen's predictive least squares (PLS) criterion can be decomposed as a sum of two terms; one measures the goodness of fit and the other penalizes the complexity of the selected model. Using this decomposition, sufficient conditions for PLS to be strongly consistent in stochastic regression models are given, and the asymptotic equivalence between PLS and the Bayesian information criterion (BIC) is established. Moreover, a new criterion, FIC, is introduced and shown to share most asymptotic properties with PLS while removing some of the difficulties encountered by PLS in finite-sample situations. In [38], the first complete proof of an analogous property for Akaike's information criterion (AIC) in determining the order of a vector autoregressive model used to fit a weakly stationary time series is given, while in [41], AIC is shown to be asymptotically efficient for same-realization predictions. Closely related papers on model selection and adaptive prediction are [39], [42] and [43].

E. Probability theory, stochastic processes and other topics

In [4] and [5], sufficient conditions are given for the law of the iterated logarithm to hold for random subsequences, least squares estimates in linear regression models and partial sums of linear processes. Papers [8] and [14] provide sufficient conditions for a general linear process to be a convergence system, while [10] considers martingale difference sequences that satisfy a local Marcinkiewicz-Zygmund condition. Papers [24], [25] and [26] resolve long-standing estimation problems in branching processes with immigration. Paper [35] studies the asymptotic behavior of the residual empirical process in stochastic regression models. In [36], uniform convergence of sample second moments is established for families of time series arrays, whose modeling by multistep prediction or likelihood methods is considered in [40]. Paper [11], [29], [30] and [33] investigate moment inequalities and their statistical applications. Density estimation, mixtures, weak convergence of recursions and sequential analysis are considered in [31], [32], [34] and [37].

Publications of Ching-Zong Wei

[1] Strong consistency of least squares estimates in multiple regression. *Proc. Nat. Acad. Sci. USA* **75** (1978), 3034–3036. (With T. L. Lai and H. Robbins.)

[2] Strong consistency of least squares estimates in multiple regression II. *J. Multivariate Anal.* **9** (1979), 343–462. (With T. L. Lai and H. Robbins.)

[3] Convergence systems and strong consistency of least squares estimates in regression models. *J. Multivariate Anal.* **11** (1981), 319–333. (With G. J. Chen and T. L. Lai.)

[4] Iterated logarithm laws with random subsequences. *Z. Warsch. verw. Gebiete* **57** (1981), 235–251. (With Y. S. Chow, H. Teicher and K. F. Yu.)
[5] A law of the iterated logarithm for double arrays of independent random variables with applications to regression and series models. *Ann. Probab.* **10** (1982), 320–335. (With T. L. Lai.)
[6] Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10** (1982), 154–166. (With T. L. Lai.)
[7] Asymptotic properties of projections with applications to stochastic regression problems. *J. Multivariate Anal.* **12** (1982), 346–370. (With T. L. Lai.)
[8] Lacunary systems and generalized linear processes. *Stoch. Process. Appl.* **14** (1983), 187–199. (With T. L. Lai.)
[9] Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. *J. Multivariate Anal.* **13** (1982), 1–23. (With T. L. Lai.)
[10] A note on martingale difference sequences satisfying the local Marcinkiewicz-Zygmund condition. *Bull. Inst. Math. Acad. Sinica* **11** (1983), 1–13. (With T. L. Lai.)
[11] Moment inequalities with applications to regression and time series models. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.), 165–172. Monograph Series, Institute of Mathematical Statistics, 1984. (With T. L. Lai.)
[12] Asymptotic properties of multivariate weighted sums with application to stochastic regression in linear dynamic systems. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.), 373–393. North-Holland, Amsterdam, 1985. (With T. L. Lai.)
[13] Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Trans. Auto. Contr.* **30** (1985), 330–338. (With A. Becker and P. R. Kumar.)
[14] Orthonormal Banach systems with applications to linear processes. *Z. Warsch. verw. Gebiete* **70** (1985), 381–393. (With T. L. Lai.)
[15] Asymptotic properties of least squares estimates in stochastic regression models. *Ann. Statist.* **13** (1985), 1498–1508.
[16] On the concept of excitation in least squares identification and adaptive control. *Stochastics* **16** (1986), 227–254. (With T. L. Lai.)
[17] Extended least squares and their application to adaptive control and prediction in linear systems. *IEEE Trans. Auto Contr.* **31** (1986), 898–906. (With T. L. Lai.)
[18] Asymptotically efficient self-tuning regulators. *SIAM J. Contr. Optimization* **25** (1987), 466–481. (With T. L. Lai.)
[19] Asymptotic inference for nearly nonstationary AR(1) process. *Ann. Statist.*, **15** (1987), 1050–1063. (With N. H. Chan.)
[20] Multivariate adaptive stochastic approximation. *Ann. Statist.* **15** (1987), 1115–1130.
[21] Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15** (1987), 1667–1682.
[22] Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16** (1988), 367–401. (With N. H. Chan.)

[23] Irreversible adaptive allocation rules. *Ann. Statist.* **17** (1989), 801–823. (With I. Hu.)
[24] Some asymptotic results for the branching process with immigration. *Stoch. Process. Appl.* **31** (1989), 261–282. (With J. Winnicki.)
[25] Estimation of the means in the branching process with immigration. *Ann. Statist.* **18** (1990), 1757–1778. (With J. Winnicki.)
[26] Convergence rates for the critical branching process with immigration. *Statist. Sinica* **1** (1991), 175–184.
[27] On predictive least squares principles. *Ann. Statist.* **20** (1992), 1–42.
[28] Martingale transforms with non-atomic limits and stochastic approximation. *Probab. Theory Related Fields* **95** (1993), 103–114.
[29] Moment bounds for deriving time series CLT's and model selection procedures. *Statist. Sinica* **3** (1993), 453–480. (With D. F. Findley.)
[30] A lower bound for expectation of a convex functional. *Statist. Probab. Letters* **18** (1993), 191–194. (With M. H. Guo.)
[31] A regression point of view toward density estimation. *J. Nonparametric Statist.* **4** (1994), 191–201. (With C. K. Chu.)
[32] How to mix random variables. *J. Chinese Statist. Asso.* **32** (1994), 295–300.
[33] A moment inequality for products. *J. Chinese Statist. Asso.* **33** (1995), 429–436. (With Y. S. Chow.)
[34] Weak convergence of recursion. *Stoch. Process. Appl.* **68** (1997), 65–82. (With G. K. Basak and I. Hu.)
[35] On residual empirical processes of stochastic regression models with applications to time series. *Ann. Statist.* **27** (1999), 237–261. (With S. Lee.)
[36] Uniform convergence of sample second moments of families of time series arrays. *Ann. Statist.* **29** (2001), 815–838. (With D. F. Findley and B. M. Pötscher.)
[37] Comments on "Sequential Analysis: Some Classical Problems and New Challenges" by T. L. Lai. *Statist. Sinica* **11** (2001), 378–379.
[38] AIC, overfitting principles, and the boundness of moments of inverse matrices for vector autoregressions and related models. *J. Multivariate Anal.* **83** (2002), 415–450. (With D. F. Findley.)
[39] On same-realization prediction in an infinite-order autoregressive process. *J. Multivariate Anal.* **85** (2003), 130–155. (With C. K. Ing.)
[40] Modeling of time series arrays by multistep prediction or likelihood methods. *J. Econometrics* **118** (2004), 151–187. (With D. F. Findley and B. M. Pötscher.)
[41] Order selection for the same-realization prediction in autoregressive processes. *Ann. Statist.* **33** (2005), 2423–2474. (With C. K. Ing.)
[42] A maximal moment inequality for long range dependent time series with applications to estimation and model selection. *Statist. Sinica* **16** (2006), 721–740. (With C. K. Ing.)
[43] Forecasting unstable processes. In *Time Series and Related Topics* (H. C. Ho, C. K. Ing and T. L. Lai, eds.). Monograph Series, Institute of Mathematical Statistics, 2006. (With J. L. Lin.)

Ching-Zong Wei, Maryland 1985.

In Hualian, Taiwan, with wife and daughter, 2004.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 1–19

DOI: 10.1214/074921706000000923

Pile-up probabilities for the Laplace likelihood estimator of a non-invertible first order moving average

F. Jay Breidt[1,*,†], Richard A. Davis[1,†,‡], Nan-Jung Hsu [2] and Murray Rosenblatt [3]

Colorado State University, National Tsing-Hua University and University of California at San Diego

Abstract: The first-order moving average model or MA(1) is given by $X_t = Z_t - \theta_0 Z_{t-1}$, with independent and identically distributed $\{Z_t\}$. This is arguably the simplest time series model that one can write down. The MA(1) with unit root ($\theta_0 = 1$) arises naturally in a variety of time series applications. For example, if an underlying time series consists of a linear trend plus white noise errors, then the differenced series is an MA(1) with unit root. In such cases, testing for a unit root of the differenced series is equivalent to testing the adequacy of the trend plus noise model. The unit root problem also arises naturally in a signal plus noise model in which the signal is modeled as a random walk. The differenced series follows a MA(1) model and has a unit root if and only if the random walk signal is in fact a constant.

The asymptotic theory of various estimators based on Gaussian likelihood has been developed for the unit root case and nearly unit root case ($\theta = 1+\beta/n, \beta \leq 0$). Unlike standard $1/\sqrt{n}$-asymptotics, these estimation procedures have $1/n$-asymptotics and a so-called pile-up effect, in which $\mathrm{P}(\hat{\theta} = 1)$ converges to a positive value. One explanation for this pile-up phenomenon is the lack of identifiability of θ in the Gaussian case. That is, the Gaussian likelihood has the same value for the two sets of parameter values (θ, σ^2) and $(1/\theta, \theta^2\sigma^2)$. It follows that $\theta = 1$ is always a critical point of the likelihood function. In contrast, for non-Gaussian noise, θ is identifiable for all real values. Hence it is no longer clear whether or not the same pile-up phenomenon will persist in the non-Gaussian case. In this paper, we focus on limiting pile-up probabilities for estimates of θ_0 based on a Laplace likelihood. In some cases, these estimates can be viewed as Least Absolute Deviation (LAD) estimates. Simulation results illustrate the limit theory.

1. Introduction

The moving average model of order one (MA(1)) given by

$$X_t = Z_t - \theta_0 Z_{t-1}, \tag{1.1}$$

[1]Department of Statistics, Colorado State University, Ft. Collins, CO 80523, USA, e-mail: jbreidt@stat.colostate.edu; rdavis@stat.colostate.edu

[2]Institute of Statistics, National Tsing-Hua University, Hsinchu, Taiwan, e-mail: njhsu@stat.nthu.edu.tw

[3]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA, e-mail: mrosenblatt@ucsd.edu

*Research supported by NSF grant DMS-9972015.

†Research supported by EPA STAR grant CR-829095.

‡Research supported by NSF grant DMS-0308109.

AMS 2000 subject classifications: primary 62M10; secondary 60F05.

Keywords and phrases: noninvertible moving averages, Laplace likelihood.

where $\{Z_t\}$ is a sequence of independent and identically distributed random variables with mean 0 and variance σ^2, is one of the simplest models in time series. The MA(1) model is invertible if and only if $|\theta_0| < 1$, since in this case Z_t can be represented explicitly in terms of past values of the X_t, i.e.,

$$Z_t = \sum_{j=0}^{\infty} \theta_0^j X_{t-j}.$$

Under this invertibility constraint, standard estimation procedures that produce asymptotically normal estimates are readily available. For example, if $\hat{\theta}$ represents the maximum likelihood estimator, found by maximizing the Gaussian likelihood based on the data $X_1, \ldots, X_n$, then it is well known (see Brockwell and Davis [3]), that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, 1 - \theta_0^2)\,. \tag{1.2}$$

From the form of the limiting variance in (1.2), the asymptotic behavior of $\hat{\theta}$, let alone the scaling, is not immediately clear in the unit root case corresponding to $\theta_0 = 1$.

In the Gaussian case, the parameters θ_0 and σ^2 are not identifiable without the constraint $|\theta_0| \leq 1$. In particular, the profile Gaussian log-likelihood, obtained by concentrating out the variance parameter, satisfies

$$L(\theta) = L(1/\theta)\,.$$

It follows that $\theta = 1$ is a critical value of the profile likelihood and hence there is a positive probability that $\theta = 1$ is indeed the maximum likelihood estimator. If $\theta_0 = 1$, then it turns out that this probability does not vanish asymptotically (see for example Anderson and Takemura [1], Tanaka [7], and Davis and Dunsmuir [6]). This phenomenon is referred to as the pile-up effect. For the case that $\theta_0 = 1$ or is near one in the sense that $\theta_0 = 1 + \gamma/n$, it was shown in Davis and Dunsmuir [6] that

$$n(\hat{\theta} - \theta_0) \xrightarrow{d} \xi_\gamma,$$

where ξ_γ is random variable with a discrete component at 0, corresponding to the asymptotic pile-up probability, and a continuous component on $(-\infty, 0)$.

The MA(1) with unit root ($\theta_0 = 1$) arises naturally in a variety of time series applications. For example, if an underlying time series consists of a linear trend plus white noise errors, then the differenced series is an MA(1) with a unit root. In such cases, testing for a unit root of the differenced series is equivalent to testing the adequacy of the trend plus noise model. The unit root problem also arises naturally in a signal plus noise model in which the signal is modeled as a random walk. The differenced series follows a MA(1) model and has a unit root if and only if the random walk signal is in fact a constant.

For Gaussian likelihood estimation, the pile-up effect is directly attributable to the non-identifiability of θ_0 in the unconstrained parameter space. On the other hand, if the data are non-Gaussian, then θ_0 is identifiable (see Breidt and Davis [2]). In this paper, we focus on the pile-up probability for estimates based on a Laplace likelihood. Assuming a Laplace distribution for the noise, we derive an expression for the joint likelihood of θ and z_{init}, where z_{init} is an augmented variable that is treated as a parameter and the scale parameter σ is concentrated out of the likelihood. If z_{init} is set equal to 0, then the resulting joint likelihood corresponds

to the least absolute deviation (LAD) objective function and the estimator of θ is referred to as the LAD estimator of θ_0. The exact likelihood can be obtained by integrating out z_{init}. In this case the resulting estimator is referred to as the quasi-maximum likelihood estimator of θ_0. It turns out that the estimator based on maximizing the joint likelihood always has a positive pile-up probability in the limit regardless of the true noise distribution. In contrast, the quasi-maximum likelihood estimator has a limiting pile-up probability of zero.

In Section 2, we describe the main asymptotic results. We begin by deriving an expression for computing the joint likelihood function based on the observed data and the augmented variable Z_{init}, in terms of the density function of the noise. The exact likelihood function can then be computed by integrating out Z_{init}. After a reparameterizion, we derive the limiting behavior of the joint likelihood for the case when the noise is assumed to follow a Laplace distribution. In Section 3, we focus on the problem of calculating asymptotic pile-up probabilities for estimators which minimize the joint Laplace likelihood (as a function of θ and z_{init}) and the exact Laplace likelihood. Section 4 contains simulation results which illustrate the asymptotic theory of Section 3.

2. Main result

Let $\{X_t\}$ be the MA(1) model given in (1.1) where $\theta_0 \in \mathbb{R}$, $\{Z_t\}$ is a sequence of iid random variables with $EZ_t = 0$ and density function f_Z. In order to compute the likelihood based on the observed data $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$, it is convenient to define an augmented initial variable Z_{init} defined by

$$Z_{init} = \begin{cases} Z_0, & \text{if } |\theta| \le 1, \\ Z_n - \sum_{t=1}^n X_t, & \text{otherwise.} \end{cases}$$

A straightforward calculation shows that the joint density of the observed data $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)'$ and the initial variable Z_{init} satisfies

$$f_{\boldsymbol{X}, Z_{init}}(\boldsymbol{x}_n, z_{init}) = \prod_{j=0}^{n} f_Z(z_j) \left(1_{\{|\theta| \le 1\}} + |\theta|^{-n} 1_{\{|\theta|>1\}}\right),$$

where the *residuals* $\{z_t\}$ are functions of $\boldsymbol{X}_n = \boldsymbol{x}_n$, θ, and $Z_{init} = z_{init}$ which can be solved forward by $z_t = X_t + \theta z_{t-1}$ for $t = 1, 2, \ldots, n$ with the initial $z_0 = z_{init}$ if $|\theta| \le 1$ and backward by $z_{t-1} = \theta^{-1}(z_t - X_t)$ for $t = n, n-1, \ldots, 1$ with the initial $z_n = z_{init} + \sum_{t=1}^n X_t$, if $|\theta| > 1$.

The Laplace log-likelihood is obtained by taking the density function for Z_t to be $f_Z(z) = \exp\{-|z|/\sigma\}/(2\sigma)$. If we view z_{init} as a parameter, then the *joint* log-likelihood is given by

$$-(n+1)\log 2\sigma - \frac{1}{\sigma}\sum_{t=0}^{n} |z_t| - n(\log|\theta|)1_{\{|\theta|>1\}}\,. \tag{2.1}$$

Maximizing this function with respect to the scale parameter σ, we obtain

$$\hat{\sigma} = \sum_{t=0}^{n} |z_t|/(n+1).$$

It follows that maximizing the joint Laplace log-likelihood is equivalent to minimizing the following objective function,

$$\ell_n(\theta, z_{init}) = \begin{cases} \sum_{t=0}^{n} |z_t|, & \text{if } |\theta| \le 1, \\ \sum_{t=0}^{n} |z_t||\theta|, & \text{otherwise.} \end{cases} \tag{2.2}$$

In order to study the asymptotic properties of the minimizer of ℓ_n when the model $\theta_0 = 1$, we follow Davis and Dunsmuir [6] by building the sample size into the parameterization of θ. Specifically, we use

$$\theta = 1 + \frac{\beta}{n}, \tag{2.3}$$

where β is any real number. Additionally, since we are also treating z_{init} as a parameter, this term is reparameterized as

$$z_{init} = Z_0 + \frac{\alpha\sigma}{\sqrt{n}}. \tag{2.4}$$

Under the (β, α) parameterization, minimizing ℓ_n with respect to θ and z_{init} is equivalent to minimizing the function,

$$U_n(\beta, \alpha) \equiv \frac{1}{\sigma} \left[\ell_n(\theta, z_{init}) - \ell_n(1, Z_0) \right],$$

with respect to β and α. The following theorem describes the limiting behavior of U_n.

Theorem 2.1. *For the model* (1.1) *with* $\theta_0 = 1$, *assume the noise sequence* $\{Z_t\}$ *is IID with* $EZ_t = 0$, $E[\ sign(Z_t)] = 0$ *(i.e., median of* Z_t *is zero),* $EZ_t^4 < \infty$ *and common probability density function* $f_Z(z) = \sigma^{-1} f(z/\sigma)$, *where* $\sigma > 0$ *is the scale parameter. We further assume that the density function* f_Z *has been normalized so that* $\sigma = E|Z_t|$. *Then*

$$U_n(\beta, \alpha) \xrightarrow{fidi} U(\beta, \alpha), \tag{2.5}$$

where $\xrightarrow{fidi}$ *denotes convergence in distribution of finite dimensional distributions and*

$$\begin{aligned} U(\beta, \alpha) = &\int_0^1 \left[\beta \int_0^s e^{\beta(s-t)} dS(t) + \alpha e^{\beta s} \right] dW(s) \\ &+ f(0) \int_0^1 \left[\beta \int_0^s e^{\beta(s-t)} dS(t) + \alpha e^{\beta s} \right]^2 ds, \end{aligned} \tag{2.6}$$

for $\beta \le 0$, *and*

$$\begin{aligned} U(\beta, \alpha) = &\int_0^1 \left[-\beta \int_{s+}^1 e^{-\beta(t-s)} dS(t) + \alpha e^{-\beta(1-s)} \right] dW(s) \\ &+ f(0) \int_0^1 \left[-\beta \int_s^1 e^{-\beta(t-s)} dS(t) + \alpha e^{-\beta(1-s)} \right]^2 ds, \end{aligned} \tag{2.7}$$

for $\beta > 0$, *in which* $S(t)$ *and* $W(t)$ *are the limits of the following partial sums*

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{[nt]} Z_i/\sigma, \quad W_n(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{[nt]} sign(Z_i),$$

respectively.

Remark. The stochastic integrals in (2.6) and (2.7) refer to Itô integrals. The double stochastic stochastic integral in the first term on the right side of (2.7) is computed as

$$\int_0^1 \int_{s+}^1 e^{-\beta(t-s)} dS(t) dW(s) = \int_0^1 e^{-\beta t} dS(t) \int_0^1 e^{\beta s} dW(s) - \int_0^1 \int_0^s e^{-\beta(t-s)} dS(t) dW(s) - \int_0^1 dS(t) dW(t),$$

where (see (2.15) below)

$$\int_0^1 dS(t) dW(t) = E(Z_i sign(Z_i))/\sigma = E|Z_i|/\sigma = 1\,.$$

Proof. We only prove the result (2.5) for a fixed (β, α); the extension to a finite collection of (β, α)'s is relatively straightforward. First consider the case $\beta \leq 0$. For calculating the Laplace likelihood $\ell_n(\theta, z_{init})$ based on model (1.1), the residuals are solved by $z_t = X_t + \theta z_{t-1}$ for $t = 1, 2, \ldots, n$ with the initial value $z_0 = z_{init}$. Since $X_t = Z_t - Z_{t-1}$, all of the true innovations can be solved forward by $Z_t = X_t + Z_{t-1}$ for $t = 1, 2, \ldots, n$ with the initial Z_0. Therefore, the centered term $\ell_n(1, Z_0)$ can be written as

$$\ell_n(1, Z_0) = |Z_0| + \sum_{i=1}^n |X_i + X_{i-1} + \cdots + X_1 + Z_0| = \sum_{i=0}^n |Z_i|.$$

For $\beta \leq 0$, i.e., $\theta \leq 1$,

$$\begin{aligned} z_i &= X_i + \theta X_{i-1} + \cdots + \theta^{i-1} X_1 + \theta^i z_{init} \\ &= (Z_i - Z_{i-1}) + \theta(Z_{i-1} - Z_{i-2}) + \cdots + \theta^{i-1}(Z_1 - Z_0) + \theta^i z_{init} \\ &= Z_i - (1-\theta)Z_{i-1} - \theta(1-\theta)Z_{i-2} - \cdots - \theta^{i-1}(1-\theta)Z_0 - \theta^i(Z_0 - z_{init}), \end{aligned}$$

which, under the true model $\theta = 1$, implies

$$\begin{aligned} \frac{1}{\sigma}\left[\ell_n(\theta, z_{init}) - \ell_n(1, Z_0)\right] &= \frac{1}{\sigma}\left(\sum_{i=0}^n |z_i| - \sum_{i=0}^n |Z_i|\right) \\ &= \frac{1}{\sigma}\sum_{i=0}^n \left(|Z_i - y_i| - |Z_i|\right), \end{aligned} \tag{2.8}$$

where $y_0 \equiv Z_0 - z_{init}$ and

$$y_i \equiv (1-\theta)\sum_{j=0}^{i-1} \theta^{i-1-j} Z_j + \theta^i (Z_0 - z_{init}),$$

for $i = 1, 2, \ldots, n$. Using the identity

$$|Z - y| - |Z| = -y\ \mathrm{sign}(Z) + 2(y - Z)\left(1_{\{0<Z<y\}} - 1_{\{y<Z<0\}}\right) \tag{2.9}$$

for $Z \neq 0$, the equation (2.8) is expressed as two summations, the first of which is

$$
\begin{aligned}
-\sum_{i=0}^{n} \frac{y_i}{\sigma}\ \mathrm{sign}(Z_i) &= (\theta-1)\sum_{i=1}^{n}\left(\sum_{j=0}^{i-1}\theta^{i-1-j}\frac{Z_j}{\sigma}\right)\ \mathrm{sign}(Z_i) \\
&\quad + \frac{z_{init}-Z_0}{\sigma}\sum_{i=0}^{n}\theta^i\ \mathrm{sign}(Z_i) \\
&= \frac{\beta}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{i-1}\left(1+\frac{\beta}{n}\right)^{i-j-1}\frac{Z_j}{\sigma}\right]\ \mathrm{sign}(Z_i) \\
&\quad + \frac{\alpha}{\sqrt{n}}\sum_{i=0}^{n}\left(1+\frac{\beta}{n}\right)^{i}\ \mathrm{sign}(Z_i) \qquad (2.10) \\
&= \beta\int_0^1\int_0^{s-}\left(1+\frac{\beta}{n}\right)^{-nt} dS_n(t)\left(1+\frac{\beta}{n}\right)^{ns-1} dW_n(s) \\
&\quad + \alpha\int_0^1\left(1+\frac{\beta}{n}\right)^{ns} dW_n(s) \\
&\to \beta\int_0^1\int_0^s e^{\beta(s-t)}dS(t)dW(s) + \alpha\int_0^1 e^{\beta s}dW(s)\,,
\end{aligned}
$$

where the limit in (2.10) follows from a simple adaptation of Theorem 2.4 (ii) in Chan and Wei [4].

To handle the second summation in computing $U_n(\beta,\alpha)$, we approximate the sum

$$\sum_{i=0}^{n} 2\,\frac{y_i - Z_i}{\sigma}\left(1_{\{0<Z_i<y_i\}} - 1_{\{y_i<Z_i<0\}}\right)$$

by

$$\sum_{i=0}^{n} 2E\left[\frac{y_i - Z_i}{\sigma}\left(1_{\{0<Z_i<y_i\}} - 1_{\{y_i<Z_i<0\}}\right)|\mathcal{F}_{i-1}\right],$$

where $\mathcal{F}_i$ is the σ-field generated by $\{Z_j : j = 0, 1, \ldots, i\}$. First we establish convergence of the latter sum and then show that the variance of the difference in sums converges to zero. Since

$$\max_{1\le i\le n}|y_i| \to 0,$$

$y_i \in \mathcal{F}_{i-1}$, we have

$$
\begin{aligned}
2E\left[\left(\frac{y_i - Z_i}{\sigma}\right)1_{\{0<Z_i<y_i\}}|\mathcal{F}_{i-1}\right] &= 2\int_0^{y_i}\left(\frac{y_i - Z}{\sigma}\right)\frac{1}{\sigma}f(\frac{z}{\sigma})dz \\
&\approx f(0)\int_0^{y_i} 2\left(\frac{y_i - z}{\sigma}\right)d\left(\frac{z}{\sigma}\right) \\
&= f(0)\left(\frac{y_i}{\sigma}\right)^2,
\end{aligned}
$$

for $y_i > 0$, and

$$\begin{aligned} 2E\left[\left(\frac{y_i - Z_i}{\sigma}\right) 1_{\{y_i<Z_i<0\}} | \mathcal{F}_{i-1}\right] &= 2\int_{y_i}^{0} \left(\frac{y_i - z}{\sigma}\right) \frac{1}{\sigma} f(\frac{z}{\sigma}) dz \\ &\approx f(0) \int_{y_i}^{0} 2\left(\frac{y_i - z}{\sigma}\right) d\left(\frac{z}{\sigma}\right) \\ &= -f(0)\left(\frac{y_i}{\sigma}\right)^2, \end{aligned}$$

for $y_i < 0$. Combining these two cases, we have

$$2\sum_{i=0}^{n} E\left[\frac{y_i - Z_i}{\sigma}\left(1_{\{0<Z_i<y_i\}} - 1_{\{y_i<Z_i<0\}}\right) | \mathcal{F}_{i-1}\right] \approx f(0) \sum_{i=0}^{n} \left(\frac{y_i}{\sigma}\right)^2,$$

where

$$\begin{aligned} \sum_{i=0}^{n} \left(\frac{y_i}{\sigma}\right)^2 &= \sum_{i=0}^{n} \left\{ (1-\theta) \sum_{j=1}^{i-1} \theta^{i-1-j} \frac{Z_j}{\sigma} + \theta^i \frac{Z_0 - z_0}{\sigma} \right\}^2 \\ &= \sum_{i=1}^{n} \left[\frac{-\beta}{n} \sum_{j=1}^{i-1} \left(1 + \frac{\beta}{n}\right)^{i-1-j} \frac{Z_j}{\sigma} - \frac{\alpha}{\sqrt{n}} \left(1 + \frac{\beta}{n}\right)^i \right]^2 \\ &= \sum_{i=1}^{n} \left[\beta \int_0^{(i-1)/n} \left(1 + \frac{\beta}{n}\right)^{i-1-sn} dS_n(s) + \alpha \left(1 + \frac{\beta}{n}\right)^i \right]^2 \frac{1}{n} \\ &\to \int_0^1 \left[\beta \int_0^s e^{\beta(s-t)} dS(t) + \alpha e^{\beta s} \right]^2 ds \end{aligned} \tag{2.11}$$

in distribution as $n \to \infty$.

It is left to show that

$$\begin{aligned} &2\sum_{i=0}^{n} \frac{y_i - Z_i}{\sigma}\left(1_{\{0<Z_i<y_i\}} - 1_{\{y_i<Z_i<0\}}\right) \\ &\quad - 2\sum_{i=0}^{n} E\left[\frac{y_i - Z_i}{\sigma}\left(1_{\{0<Z_i<y_i\}} - 1_{\{y_i<Z_i<0\}}\right) | \mathcal{F}_{i-1}\right] \end{aligned} \tag{2.12}$$

converges to zero in probability. Define

$$y_i^* \equiv 2\,\frac{y_i - Z_i}{\sigma}\left(1_{\{0<Z_i<y_i\}} - 1_{\{y_i<Z_i<0\}}\right).$$

The expectation of (2.12) is zero and therefore, it is enough to show that the

variance of (2.12) also converges to zero. The variance of (2.12) is equal to

$$\begin{aligned}
&\sum_{i=0}^{n} \operatorname{var}\left(y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right)\right) + 2\sum_{i<j} \operatorname{cov}\left(y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right), y_j^* - E\left(y_j^*|\mathcal{F}_{j-1}\right)\right) \\
&\quad = \sum_{i=0}^{n} E\left[y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right)\right]^2 \\
&\quad = \sum_{i=0}^{n} EE\left[(y_i^*)^2 - \left(E\left(y_i^*|\mathcal{F}_{i-1}\right)\right)^2 |\mathcal{F}_{i-1}\right] \\
&\quad = \sum_{i=0}^{n} E\left[E\left((y_i^*)^2|\mathcal{F}_{i-1}\right) - \left(E\left(y_i^*|\mathcal{F}_{i-1}\right)\right)^2\right] \qquad (2.13) \\
&\quad \approx \sum_{i=0}^{n} E\left[\frac{4}{3}f(0)\left(\frac{y_i}{\sigma}\right)^3 - f(0)^2\left(\frac{y_i}{\sigma}\right)^4\right] \\
&\quad \approx \frac{4}{3}f(0)E\left[\sum_{i=0}^{n}\left(\frac{y_i}{\sigma}\right)^3\right] - f(0)^2 E\left[\sum_{i=0}^{n}\left(\frac{y_i}{\sigma}\right)^4\right] \\
&\quad \to 0,
\end{aligned}$$

as $n \to \infty$, where

$$\begin{aligned}
&\operatorname{cov}\left(y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right), y_j^* - E\left(y_j^*|\mathcal{F}_{j-1}\right)\right) \\
&\quad = E\left[y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right)\right]\left[y_j^* - E\left(y_j^*|\mathcal{F}_{j-1}\right)\right] \\
&\quad = EE\left[\left(y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right)\right)\left(y_j^* - E\left(y_j^*|\mathcal{F}_{j-1}\right)\right)\middle|\mathcal{F}_{j-1}\right] \\
&\quad = E\left[\left(y_i^* - E\left(y_i^*|\mathcal{F}_{i-1}\right)\right)E\left(y_j^* - E\left(y_j^*|\mathcal{F}_{j-1}\right)\middle|\mathcal{F}_{j-1}\right)\right] \\
&\quad = 0,
\end{aligned}$$

for $i < j$, and

$$\begin{aligned}
E\left(y_i^*|\mathcal{F}_{i-1}\right) &\approx f(0)\left(\frac{y_i}{\sigma}\right)^2, \\
E\left((y_i^*)^2|\mathcal{F}_{i-1}\right) &\approx \frac{4}{3}f(0)\left(\frac{y_i}{\sigma}\right)^3, \\
\sqrt{n}\sum_{i=0}^{n}\left(\frac{y_i}{\sigma}\right)^3 &\to -\int_0^1\left(\beta\int_0^s e^{\beta(s-t)}dS(t) + \alpha e^{\beta s}\right)^3 ds, \\
n\sum_{i=0}^{n}\left(\frac{y_i}{\sigma}\right)^4 &\to \int_0^1\left(\beta\int_0^s e^{\beta(s-t)}dS(t) + \alpha e^{\beta s}\right)^4 ds.
\end{aligned}$$

Based on (2.10), (2.11), and (2.13), the proof for $\beta \le 0$ is complete.

The proof for $\beta \ge 0$ given in (2.7) is similar to that for $\beta \le 0$. For $\beta \ge 0$, i.e., $\theta \ge 1$, the residuals $\{z_t\}$ are solved backward by $z_{t-1} = \theta^{-1}(z_t - X_t)$ for $t = n, n-1, \ldots, 1$ with the initial $z_n \equiv z_{init} + \sum_{t=1}^{n} X_t$. Solving these equations, we have

$$z_{n-1-i} = -\theta^{-1}\left(X_{n-i} + \theta^{-1}X_{n-i-1} + \cdots + \theta^{-i}X_n - \theta^{-i}z_n\right),$$

for $i = 0, 1, \ldots, n-1$. Writing $X_t = Z_t - Z_{t-1}$, we obtain

$$\begin{aligned}
-z_{n-1-i}\theta &= X_{n-i} + \theta^{-1}X_{n-i-1} + \cdots + \theta^{-i}X_n - \theta^{-i}z_n \\
&= (Z_{n-i} - Z_{n-i-1}) + \theta^{-1}(Z_{n-i+1} - Z_{n-i}) + \cdots \\
&\quad + \theta^{-i}(Z_n - Z_{n-1}) - \theta^{-i}z_n \\
&= -Z_{n-i-1} + (1-\theta^{-1})Z_{n-i} + \cdots + \theta^{-(i-1)}(1-\theta^{-1})Z_{n-1} \\
&\quad + \theta^{-i}(Z_n - z_n) \\
&= -Z_{n-i-1} + y_{n-i-1},
\end{aligned}$$

where

$$\begin{aligned}
y_{n-1-i} &\equiv \left(1-\theta^{-1}\right)\sum_{j=1}^{i}(\theta^{-1})^{i-j}Z_{n-j} + \theta^{-i}(Z_n - z_n) \\
&= \left(1-\theta^{-1}\right)\sum_{j=1}^{i}(\theta^{-1})^{i-j}Z_{n-j} + \theta^{-i}\left[\left(\sum_{i=1}^{n}X_i + Z_0\right) - \left(\sum_{i=1}^{n}X_i + z_{init}\right)\right] \\
&= \left(1-\theta^{-1}\right)\sum_{j=1}^{i}(\theta^{-1})^{i-j}Z_{n-j} + \theta^{-i}(Z_0 - z_{init}),
\end{aligned}$$

for $i = 0, 1, \ldots, n-1$ and $y_n \equiv Z_n - z_n = Z_0 - z_{init}$. Again, for $\theta \geq 1$, we have

$$\frac{1}{\sigma}\left[\ell_n(\theta, z_{init}) - \ell_n(1, Z_0)\right] = \frac{1}{\sigma}\sum_{i=0}^{n}\left(|Z_i - y_i| - |Z_i|\right),$$

which has the same form as that for $\theta \leq 1$ but with different $\{y_i\}$. Following a similar derivation for $\theta \leq 1$, one can show that

$$\begin{aligned}
-\sum_{i=1}^{n}\frac{y_i}{\sigma}\,\mathrm{sign}(Z_i) &\to -\beta\int_0^1\int_{s+}^1 e^{-\beta(t-s)}dS(t)dW(s) + \alpha\int_0^1 e^{-\beta(1-s)}dW(s), \\
\sum_{i=0}^{n}\frac{y_i^2}{\sigma^2} &\to \int_0^1\left[-\beta\int_s^1 e^{-\beta(t-s)}dS(t) + \alpha e^{-\beta(1-s)}\right]^2 ds,
\end{aligned}$$

in distribution as $n \to \infty$. Combining this with the analogous result (2.13) for $\beta \geq 0$, completes the proof. □

We close this section with some elementary results concerning the relationship between the limiting Brownian motions $S(t)$ and $W(t)$ that will be used in the sequel. Since $\sigma = E|Z_t|$, the process $S(t)$ can be decomposed as

$$S(t) = W(t) + cV(t)\,, \tag{2.14}$$

where $\{W(t)\}$ and $\{V(t)\}$ are independent standard Bronwnian motions on $[0, 1]$ and

$$c = \sqrt{\mathrm{Var}(Z_t)/\sigma^2 - 1}\,.$$

In addition, we have the following identities

$$\begin{aligned}
\int_0^1 V(s)ds &= V(1) - \int_0^1 sdV(s), \\
\int_0^1 V(s)dW(s) &= V(1)W(1) - \int_0^1 W(s)dV(s), \\
\int_0^1 dW(s)dW(s) &= \int_0^1 ds = 1, \\
\int_0^1 dV(s)dW(s) &= 0,
\end{aligned}$$

where the first two equations can be obtained easily by integration by parts. It follows that

$$\int_0^1 dS(s)dW(s) = \int_0^1 dW(s)dW(s) + c\int_0^1 dV(s)dW(s) = 1\,. \tag{2.15}$$

3. Pile-up probabilities

3.1. Joint likelihood

In this section, we will consider the local maximizer of the joint likelihood given by $-\ell_n$ in (2.2). This estimator was also studied by Davis and Dunsmuir [6] in the Gaussian case. Denote by $(\hat{\theta}_n^{(J)}, \hat{z}_{init,n}^{(J)})$ the local minimizer of $\ell_n(\theta, z_{init})$ in which $\hat{\theta}_n^{(J)}$ is closest to 1. Using the (β, α) parameterization given in (2.3) and (2.4), this is equivalent to finding the local minimizer $(\hat{\beta}_n^{(J)}, \hat{\alpha}_n^{(J)})$ of $U_n(\beta, \alpha)$ in which $\hat{\beta}_n^{(J)}$ is closest to zero. Moreover, the respective local minimizers of ℓ_n and U_n are connected through the following relations:

$$\hat{\theta}_n^{(J)} = 1 + \frac{\hat{\beta}_n^{(J)}}{n}, \quad \hat{z}_{init,n}^{(J)} = Z_0 + \frac{\hat{\alpha}_n^{(J)}\sigma}{\sqrt{n}}. \tag{3.1}$$

If the convergence of U_n to U in Theorem 1 is strengthened to weak convergence of processes on $C(\mathbb{R}^2)$, then the argument given in Davis and Dunsmuir [6] suggests the convergence in distribution of $(\hat{\beta}_n^{(J)}, \hat{\alpha}_n^{(J)})$ to $(\beta^{(J)}, \alpha^{(J)})$, where $(\hat{\beta}^{(J)}, \hat{\alpha}^{(J)})$ is the local minimizer of $U(\beta, \alpha)$ in which $\hat{\beta}^{(J)}$ is closest to 0. It follows that

$$(n(\hat{\theta}_n^{(J)} - 1), \sqrt{n}(\hat{z}_{init,n}^{(J)} - Z_0)/\sigma) \xrightarrow{d} (\hat{\beta}^{(J)}, \hat{\alpha}^{(J)})\,. \tag{3.2}$$

The proofs of these results are the subject of on-going research and will appear in a forthcoming manuscript.

Turning to the question of pile-up probabilities, we have that 1 is a local minimizer if the derivative of the criterion function from the left is negative and the derivative from the right is positive; that is,

$$\begin{aligned}
P(\hat{\theta}_n^{(J)} = 1) &= P(\hat{\beta}_n^{(J)} = 0) \\
&= P\left[\lim_{\beta\uparrow 0} \frac{\partial}{\partial\beta} U_n\left(\beta, \hat{\alpha}_n(\beta)\right) < 0 \text{ and } \lim_{\beta\downarrow 0} \frac{\partial}{\partial\beta} U_n\left(\beta, \hat{\alpha}_n(\beta)\right) > 0\right],
\end{aligned}$$

where $\hat{\alpha}_n(\beta) = \arg\min_\alpha U_n(\beta, \alpha)$ for given β. Assuming convergence of the right- and left-hand derivatives of the process $U_n(\beta, \hat{\alpha}_n(\beta))$, we obtain

$$\lim_{n\to\infty} P(\hat{\theta}_n^{(J)} = 1) = P\left[\lim_{\beta\uparrow 0} \frac{\partial}{\partial\beta} U\left(\beta, \hat{\alpha}(\beta)\right) < 0 \text{ and } \lim_{\beta\downarrow 0} \frac{\partial}{\partial\beta} U\left(\beta, \hat{\alpha}(\beta)\right) > 0\right], \tag{3.3}$$

where $\hat{\alpha}(\beta) = \arg\min_\alpha U(\beta, \alpha)$. We now proceed to simplify the limits of the two derivatives in the brackets of (3.3) in terms of the processes $S(t)$ and $W(t)$. According to (2.6) in Theorem 2.1, we have

$$\begin{aligned}
\lim_{\beta\uparrow 0} \frac{\partial}{\partial\alpha} U(\beta, \alpha) &= \lim_{\beta\uparrow 0} \left\{ \int_0^1 e^{\beta s} dW(s) + f(0) 2\alpha \int_0^1 e^{2\beta s} ds \right\} \\
&= \int_0^1 dW(s) + 2\alpha f(0) \int_0^1 ds \\
&= W(1) + 2\alpha f(0),
\end{aligned}$$

and therefore

$$\hat{\alpha}(0-) = -\frac{W(1)}{2f(0)}.$$

The derivative of $U(\beta, \alpha)$ with respect to β at zero from the left-hand side satisfies

$$\begin{aligned}
\frac{\partial}{\partial\beta} U(\beta, \alpha) = &\int_0^1 \int_0^s e^{\beta(s-t)} dS(t) dW(s) + \beta \int_0^1 \int_0^s e^{\beta(s-t)} (s-t) dS(t) dW(s) \\
&+ \alpha \int_0^1 e^{\beta s} s dW(s) \\
&+ f(0) \left\{ 2\beta \int_0^1 \left(\int_0^s e^{\beta(s-t)} dS(t) \right)^2 ds \right. \\
&\quad + \beta^2 \int_0^1 2 \left(\int_0^s e^{\beta(s-t)} dS(t) \right) \left(\int_0^s e^{\beta(s-t)} (s-t) dS(t) \right) ds \\
&\quad + \alpha^2 \int_0^1 e^{2\beta s} 2s ds + 2\alpha \int_0^1 e^{\beta s} \left(\int_0^s e^{\beta(s-t)} dS(t) \right) ds \\
&\quad \left. + 2\alpha\beta \int_0^1 e^{\beta s} \left(\int_0^s e^{\beta(s-t)} (2s-t) dS(t) \right) ds \right\}.
\end{aligned}$$

Taking the limit as $\beta \uparrow 0$, we have

$$\begin{aligned}
\lim_{\beta\uparrow 0} \frac{\partial}{\partial\beta} U(\beta, \hat{\alpha}(\beta)) = &\int_0^1 \int_0^s dS(t) dW(s) + \hat{\alpha}(0-) \int_0^1 s dW(s) \\
&+ f(0) \left\{ \hat{\alpha}^2(0-) \int_0^1 2s ds + 2\hat{\alpha}(0-) \int_0^1 \int_0^s dS(t) ds \right\} \\
= &\int_0^1 S(s) dW(s) - W(1) \int_0^1 S(s) ds \\
&+ \frac{W(1)}{2f(0)} \left[\int_0^1 W(s) ds - \frac{W(1)}{2} \right] \\
=: &\, Y.
\end{aligned} \tag{3.4}$$

Similarly, according to (2.7) in Theorem 2.1, we have

$$\begin{aligned}\lim_{\beta\downarrow 0}\frac{\partial}{\partial\alpha}U(\beta,\alpha) &= \lim_{\beta\downarrow 0}\left\{\int_0^1 e^{-\beta(1-s)}dW(s) + f(0)2\alpha\int_0^1 e^{-2\beta(1-s)}ds\right\}\\ &= \int_0^1 dW(s) + 2\alpha f(0)\int_0^1 ds\\ &= W(1) + 2\alpha f(0),\end{aligned}$$

and therefore

$$\hat{\alpha}(0+) = -\frac{W(1)}{2f(0)},$$

which is same as $\hat{\alpha}(0-)$. The derivative of $U(\beta,\alpha)$ with respect to β at zero from righthand side satisfies

$$\begin{aligned}\frac{\partial}{\partial\beta}U(\beta,\alpha) = &-\int_0^1\int_{s+}^1 e^{-\beta(t-s)}dS(t)dW(s) - \beta\int_0^1\int_s^1 e^{-\beta(t-s)}(s-t)dS(t)dW(s)\\ &+\alpha\int_0^1 e^{-\beta(1-s)}(s-1)dW(s)\\ &+f(0)\left\{2\beta\int_0^1\left(\int_s^1 e^{-\beta(t-s)}dS(t)\right)^2 ds\right.\\ &\quad+\beta^2\int_0^1 2\left(\int_s^1 e^{-\beta(t-s)}dS(t)\right)\\ &\qquad\times\left(\int_s^1 e^{-\beta(t-s)}(s-t)dS(t)\right)ds\\ &\quad+\alpha^2\int_0^1 e^{-2\beta(1-s)}2(s-1)ds\\ &\quad-2\alpha\int_0^1 e^{-\beta(1-s)}\left(\int_s^1 e^{-\beta(t-s)}dS(t)\right)ds\\ &\quad\left.-2\alpha\beta\int_0^1\int_s^1 e^{-\beta(1+t-2s)}(2s-t-1)dS(t)ds\right\}.\end{aligned}$$

Taking the limit $\beta\downarrow 0$ and using the remark in Section 2, we have

$$\begin{aligned}&\lim_{\beta\downarrow 0}\frac{\partial}{\partial\beta}U(\beta,\hat{\alpha}(\beta))\\ &\quad\to -\int_0^1\int_{s+}^1 dS(t)dW(s) + \hat{\alpha}(0+)\int_0^1(s-1)dW(s)\\ &\qquad+f(0)\left\{\hat{\alpha}^2(0+)\int_0^1 2(s-1)ds - 2\hat{\alpha}(0+)\int_0^1\int_s^1 dS(t)ds\right\}\\ &\quad= -S(1)W(1) + \int_0^1 S(s)dW(s) + 1 + \hat{\alpha}(0+)\left[[(s-1)W(s)]_0^1 - \int_0^1 W(s)ds\right]\\ &\qquad+f(0)\left\{-\hat{\alpha}^2(0+) - 2\hat{\alpha}(0+)\left[S(1) - \int_0^1 S(s)ds\right]\right\}\\ &\quad= \int_0^1 S(s)dW(s) - W(1)\int_0^1 S(s)ds + \frac{W(1)}{2f(0)}\left[\int_0^1 W(s)ds - \frac{W(1)}{2}\right] + 1\\ &\quad= Y + 1.\end{aligned}$$

Therefore, the pile-up probability in (3.3) can be expressed in terms of Y as

$$\begin{aligned}\lim_{n\to\infty} P(\hat{\theta}_n^{(J)} = 1) &= P\left[Y < 0 \text{ and } Y + 1 > 0\right] \\ &= P\left[-1 < Y < 0\right].\end{aligned}$$

3.2. Exact likelihood estimation

In this section, we consider pile-up probabilities associated with the estimator that maximizes the exact Laplace likelihood. For $\theta \leq 1$, the joint density of $(\boldsymbol{x}_n, z_{init})$ satisfies

$$\begin{aligned}f(\boldsymbol{x}_n, z_{init}) = \prod_{t=0}^{n} f(z_t) &= \left(\frac{1}{2\sigma}\right)^{n+1} \exp\left(-\frac{\sum_{t=0}^{n}|z_t|}{\sigma}\right) \\ &= \left(\frac{1}{2\sigma}\right)^{n+1} \exp\left\{-\frac{[\ell_n(\theta, z_{init}) - \ell_n(1, Z_0)] + \ell_n(1, Z_0)}{\sigma}\right\} \\ &= \left(\frac{1}{2\sigma}\right)^{n+1} \exp\left(-\frac{\sum_{t=0}^{n}|Z_t|}{\sigma}\right) e^{-U_n(\beta,\alpha)}.\end{aligned}$$

Integrating out the augmented variable z_{init}, we obtain

$$\int_{-\infty}^{\infty} f(\boldsymbol{x}_n, z_{init}) dz_{init} = \left(\frac{1}{2\sigma}\right)^{n+1} \exp\left(-\frac{\sum_{t=0}^{n}|Z_t|}{\sigma}\right) \frac{\sigma}{\sqrt{n}} \int_{-\infty}^{\infty} e^{-U_n(\beta,\alpha)} d\alpha,$$

since under the parameterization (2.4), $dz_{init} = (\sigma/\sqrt{n})d\alpha$. The Laplace log-likelihood of (θ, σ) given $\boldsymbol{x}_n$ then satisfies

$$\begin{aligned}\ell_n^*(\theta, \sigma) &\equiv \log \int_{-\infty}^{\infty} f(\boldsymbol{x}_n, z_{init}) dz_{init} \\ &= -(n+1)\log(2\sigma) - \frac{\sum_{t=0}^{n}|Z_t|}{\sigma} + \log\left(\frac{\sigma}{\sqrt{n}}\right) + \log \int_{-\infty}^{\infty} e^{-U_n(\beta,\alpha)} d\alpha,\end{aligned}$$

where the last term does not depend on σ as $n \to \infty$. So maximizing ℓ_n^* with respect to $\theta \leq 1$ is approximately the same as maximizing

$$U_n^*(\beta) = \log \int_{-\infty}^{\infty} e^{-U_n(\beta,\alpha)}\, d\alpha \tag{3.5}$$

with respect to $\beta \leq 0$,

Similarly, for $\theta > 1$, the Laplace log-likelihood of (θ, σ) is

$$\begin{aligned}\ell_n^*(\theta, \sigma) &\equiv \log \int_{-\infty}^{\infty} f(\boldsymbol{x}_n, z_{init}) dz_{init} \\ &= -n\log|\theta| - (n+1)\log(2\sigma) - \frac{\sum_{t=0}^{n}|Z_t|}{\sigma|\theta|} \\ &\quad + \log\left(\frac{\sigma}{\sqrt{n}}\right) + \log \int_{-\infty}^{\infty} e^{-U_n(\beta,\alpha)|\theta|^{-1}} d\alpha,\end{aligned}$$

where again the last term does not depend on σ as $n \to \infty$. As above, maximizing ℓ_n^* with respect to $\theta > 1$ is equivalent to maximizing

$$U_n^*(\beta) = \log \int_{-\infty}^{\infty} e^{-U_n(\beta,\alpha)n/(n+\beta)}\, d\alpha \tag{3.6}$$

for $\beta > 0$.

A heuristic argument based on the process convergence of U_n to U suggests that

$$U_n^*(\beta) \to U^*(\beta) = \log \int_{-\infty}^{\infty} e^{-U(\beta,\alpha)}\, d\alpha\,, \tag{3.7}$$

where U_n^* is specified by (3.5) for $\beta \le 0$ and by (3.6) for $\beta > 0$. Now if $\hat{\beta}_n^{(E)}$ denotes the local maximum of the exact likelihood, or alternatively the maximizer of $U_n^*(\beta)$ that is closest to 0, then the convergence in (3.7) suggests convergence in distribution for the local maximizer of the exact likelihood, i.e.,

$$n(\hat{\theta}_n^{(E)} - 1) = \hat{\beta}_n^{(E)} \xrightarrow{d} \hat{\beta}^{(E)}\,, \tag{3.8}$$

where $\hat{\beta}^{(E)}$ is the local maximizer of $U^*(\beta)$ that is closest to 0.

The limiting pile-up probabilities for $\hat{\theta}_n^{(E)}$ are calculated from

$$\begin{aligned}\lim_{n\to\infty} P(\hat{\theta}_n^{(E)} = 1) &= \lim_{n\to\infty} P(\hat{\beta}_n^{(E)} = 0) = P(\hat{\beta}^{(E)} = 0) \\ &= P\left(\lim_{\beta\uparrow 0} \frac{\partial}{\partial\beta} U^*(\beta) > 0 \text{ and } \lim_{\beta\downarrow 0} \frac{\partial}{\partial\beta} U^*(\beta) < 0\right).\end{aligned}$$

Fortunately, the right- and left-hand derivatives of U^* can be computed explicitly. These are found to be

$$\begin{aligned}\lim_{\beta\uparrow 0} \frac{\partial}{\partial\beta} U^*(\beta) &= -\frac{W^2(1)}{4f(0)} + \frac{W(1)}{2f(0)} \int_0^1 W(s)ds - W(1)\int_0^1 S(s)ds + \int_0^1 S(s)dW(s) \\ &\quad + \frac{1}{2} \\ &= Y + \frac{1}{2}, \\ \lim_{\beta\downarrow 0} \frac{\partial}{\partial\beta} U^*(\beta) &= -\frac{W^2(1)}{4f(0)} + \frac{W(1)}{2f(0)} \int_0^1 W(s)ds - W(1)\int_0^1 S(s)ds + \int_0^1 S(s)dW(s) \\ &\quad + \frac{1}{2} \\ &= Y + \frac{1}{2},\end{aligned}$$

where Y is defined in (3.4). The limiting pile-up probability for $\hat{\theta}_n^{(E)}$ is then

$$\lim_{n\to\infty} P(\hat{\theta}_n^{(E)} = 1) = P\left[-\frac{1}{2} < Y < -\frac{1}{2}\right] = 0.$$

3.3. Remarks

Here we collect several remarks concerning the results of Sections 3.1 and 3.2.

Remark 1. Under the assumptions of Theorem 2.1, the asymptotic pile-up probability for estimator $\hat{\theta}_n^{(J)}$ based on the joint likelihood is always positive. On the other hand, the asymptotic pile-up probability for estimator $\hat{\theta}_n^{(E)}$ based on the exact likelihood is zero.

Remark 2. The two estimators of θ_0 considered in Sections 3.1 and 3.2 were defined as the local optimizers of objective functions that were closest to 1. One could also consider the global optimizers of these objective functions. For example, the exact MLE in the Gaussian case was considered in Davis and Dunsmuir [6] and Davis, Chen and Dunsmuir [5] and has a different limiting distribution than the local MLE. In our case, there will be a positive asymptotic pile-up probability for the global maximum of the joint likelihood and a zero asymptotic pile-up probability for the global maximum of the exact likelihood.

Remark 3. Suppose Z_t has a Laplace distribution with the density function

$$f_Z(z) = \frac{1}{2\sigma} e^{-|z|/\sigma}.$$

Then Y defined in (3.4) satisfies

$$Y = \int_0^1 [W(1)s - W(s)]\, dV(s) - \frac{1}{2}, \tag{3.9}$$

where $W(s)$ and $V(s)$ are independent standard Brownian motions. To prove (3.9), note that the constant c in (2.14) is equal to 1 so that

$$S(t) = W(t) + V(t).$$

In the following calculations, we use the well-known Itô formula

$$\int_0^1 W(s)dW(s) = \frac{W^2(1)}{2} - \frac{1}{2}.$$

Since $f(0) = 1/2$, the random variable Y defined in (3.4) can be further simplified in terms of $W(t)$ and $V(t)$ as

$$\begin{aligned}
Y &= \int_0^1 S(s)dW(s) - W(1)\int_0^1 S(s)ds + \frac{W(1)}{2f(0)}\left[\int_0^1 W(s)ds - \frac{W(1)}{2}\right] \\
&= \int_0^1 V(s)dW(s) + \int_0^1 W(s)dW(s) - W(1)\int_0^1 V(s)ds - W(1)\int_0^1 W(s)ds \\
&\quad + W(1)\int_0^1 W(s)ds - \frac{W^2(1)}{2} \\
&= V(1)W(1) - \int_0^1 W(s)dV(s) + \frac{W^2(1)}{2} - \frac{1}{2} - W(1)\left[V(1) - \int_0^1 sdV(s)\right] \\
&\quad - \frac{W^2(1)}{2} \\
&= \int_0^1 [W(1)s - W(s)]\, dV(s) - \frac{1}{2}.
\end{aligned}$$

Therefore, the pile-up probability for Laplace innovations is

$$\begin{aligned}
&P\left(-1 < Y < 0\right) \\
&\quad = P\left(-\frac{1}{2} < \int_0^1 \left[W(1)s - W(s)\right] dV(s) < \frac{1}{2}\right) \\
&\quad = E\left[P\left(-\frac{1}{2} < \int_0^1 \left[W(1)s - W(s)\right] dV(s) < \frac{1}{2}\right) \,\middle|\, W(t) \text{ on } t \in [0,1]\right] \\
&\quad = E\left[P\left(-\frac{1}{2}\left\{\int_0^1 [W(1)s - W(s)]^2 ds\right\}^{-1/2} < U \right.\right. \\
&\qquad\qquad \left.\left. < \frac{1}{2}\left\{\int_0^1 [W(1)s - W(s)]^2 ds\right\}^{-1/2}\right)\right] \\
&\quad = E\left[\Phi\left(\frac{1}{2}\left\{\int_0^1 [W(1)s - W(s)]^2 ds\right\}^{-1/2}\right)\right. \\
&\qquad \left. - \Phi\left(-\frac{1}{2}\left\{\int_0^1 [W(1)s - W(s)]^2 ds\right\}^{-1/2}\right)\right] \\
&\quad \approx 0.820,
\end{aligned}$$

where U has the standard normal distribution and $\Phi(\cdot)$ is the corresponding cumulative distribution function. This pile-up probability, which was computed via simulation based on 100000 replications of $W(t)$ on $[0,1]$, has a standard error of 0.0010.

Remark 4. From the limiting result (3.2), it follows that the random variable Z_0 can be *estimated* consistently. It may seem odd to have a consistent estimate of a noise term in a moving average process. On the other hand, an MA(1) process with a unit root is both invertible and non-invertible. That is, Z_0 is an element of the two Hilbert spaces generated by the linear span of $\{X_t, t \le 0\}$ and $\{X_t, t \ge 1\}$, respectively. It is the latter Hilbert space which allows for consistent estimation of Z_0.

4. Numerical simulation

In this section, we compute the asymptotic pile-up probabilities associated with the estimator $\hat{\theta}^{(J)}$ which maximizes the joint Laplace likelihood for several different noise distributions. The empirical properties of estimators $\hat{\theta}_n^{(J)}$ (the local maximizer of the joint Laplace likelihood) and $\hat{\theta}_n^{(E)}$ (the local maximizer of the exact Laplace likelihood) for finite samples are compared with each other and with the corresponding asymptotic theory.

For approximating the asymptotic pile-up probabilities and limiting distribution of $\hat{\beta}_n^{(J)}$, we first simulate 100000 replications of independent standard Wiener processes $W(t)$ and $V(t)$ on $[0,1]$ in which $W(t)$ and $V(t)$ are approximated by the partial sums $W(t) = \sum_{j=1}^{[10000t]} W_j/\sqrt{10000}$ and $V(t) = \sum_{j=1}^{[10000t]} V_j/\sqrt{10000}$, where $\{W_j\}$ and $\{V_j\}$ are independent standard normal random variables. From the simulation of $W(t)$ and $V(t)$, the distribution of the limit random variable $\hat{\beta}^{(J)}$ can be tabulated and the pile-up probability $P(-1 < Y < 0)$ estimated, where Y is given in (3.4). The empirical pile-up probabilities and their asymptotic limits are

displayed in Table 1 for different noise distributions: Laplace, Gaussian, uniform, and t with 5 degrees of freedom. Notice that there is good agreement between the asymptotic and empirical probabilities for sample sizes as small as 50.

For examining the empirical performance of the local maximizers $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$, we only consider the process generated with Laplace noise with $\sigma = 1$ and sample sizes $n = 20, 50, 100, 200$. For each setup, 1000 realizations of the MA(1) process with $\theta_0 = 1$ are generated and the estimates $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$ and their corresponding estimates of the scale parameter are obtained. The estimation results are summarized in Table 2. For comparison, the standard deviation based on the limit distributions of $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$ are also reported (denoted by asymp in the table), which are obtained numerically based on 100000 replicates of the limit process U. Generally speaking, the empirical root mean square errors are very close to their asymptotic values even for very small samples. Moreover, the estimation error of $\hat{\theta}_n^{(J)}$ is about 1/2 the estimation error of $\hat{\theta}_n^{(E)}$, which indicates the superiority of using the joint likelihood over exact likelihood when $\theta_0 = 1$.

We also considered performance of the two estimators $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$ in the case when $\theta_0 \neq 1$. A limit theory for these estimators can be derived in this case by assuming that the true value θ_0 is near 1. That is, we can parameterize the MA(1) parameter by $\theta_0 = 1 + \gamma/n$ (e.g., Davis and Dunsmuir [6]). While we have not pursued the theory in the near unit root case, the relative performance of these

TABLE 1

Empirical pile-up probabilities of the local maximizer $\hat{\theta}_n^{(J)}$ of the joint Laplace likelihood for an MA(1) with $\theta_0 = 1$ and sample sizes $n = 20, 50, 100, 200$ (based on 1000 replicates) and their asymptotic values under various noise distributions.

n	Gau	Lap	Unif	$t(5)$
20	0.827	0.796	0.831	0.796
50	0.859	0.806	0.864	0.823
100	0.873	0.819	0.864	0.817
200	0.844	0.819	0.843	0.831
500	0.855	0.809	0.841	0.846
∞	0.873	0.820	0.862	0.836

TABLE 2

Bias, standard deviation and root mean square error of the local maximizers $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$ of the joint and exact Laplace likelihoods, respectively, for an MA(1) process generated by Laplace noise with $\theta_0 = 1$ and $\sigma = 1$ (1000 replications).

n		$\hat{\theta}_n^{(J)}$	$\hat{\theta}_n^{(E)}$
$n = 20$	bias	-0.003	-0.006
	s.d.	0.066	0.144
	rmse	0.066	0.144
	asymp	0.053	0.121
$n = 50$	bias	-0.000	0.000
	s.d.	0.021	0.057
	rmse	0.021	0.057
	asymp	0.021	0.048
$n = 100$	bias	-0.000	0.001
	s.d.	0.011	0.030
	rmse	0.011	0.030
	asymp	0.011	0.024
$n = 200$	bias	0.000	0.001
	s.d.	0.006	0.014
	rmse	0.006	0.014
	asymp	0.005	0.012

TABLE 3

Bias, standard deviation and root mean square error of the global maximizers $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$ of the joint and exact Laplace likelihoods, respectively, for an MA(1) process generated by Laplace noise with $\theta_0 = 0.8, 0.9, 0.95, 1/0.95, 1/0.9, 1/0.8$, $\sigma = 1$, and $n = 50$ based on 1000 replications. First 2 columns record the number of times (out of 1000) that the estimates were less than 1 (invertible) and equal to 1 (unit root).

θ_0		< 1	$= 1$	bias	s.d.	rmse
0.8	$\hat{\theta}_{50}^{(J)}$	789	95	0.0734	0.1973	0.2105
	$\hat{\theta}_{50}^{(E)}$	873	19	0.0498	0.1753	0.1822
0.9	$\hat{\theta}_{50}^{(J)}$	557	322	0.0578	0.1398	0.1513
	$\hat{\theta}_{50}^{(E)}$	767	93	0.0327	0.0933	0.0989
0.95	$\hat{\theta}_{50}^{(J)}$	404	503	0.0322	0.0708	0.0778
	$\hat{\theta}_{50}^{(E)}$	632	168	0.0235	0.0821	0.0854
1/0.95	$\hat{\theta}_{50}^{(J)}$	90	540	-0.0315	0.0763	0.0825
	$\hat{\theta}_{50}^{(E)}$	286	114	-0.0207	0.0890	0.0914
1/0.9	$\hat{\theta}_{50}^{(J)}$	89	299	-0.0389	0.1227	0.1287
	$\hat{\theta}_{50}^{(E)}$	207	71	-0.0327	0.1218	0.1261
1/0.8	$\hat{\theta}_{50}^{(J)}$	96	109	-0.0338	0.2645	0.2666
	$\hat{\theta}_{50}^{(E)}$	149	19	-0.0492	0.2280	0.2333

estimators was compared in a limited simulation study. We considered 3 values of $\theta_0 = 0.8, 0.9, 0.95$ and their reciprocals $1/0.8, 1/0.9, 1/0.95$. The latter 3 cases correspond to purely non-invertible models. The results reported in Table 3 are based on the global optimization of the joint and exact likelihoods. The first two columns contain the number of realizations out of 1000 in which the estimator was invertible (< 1) and on the unit circle ($= 1$), respectively. For example, in the $\theta_0 = 0.8$ and $\hat{\theta}_n^{(J)}$ case, 78.9% of the realizations produced invertible models, and the empirical pile-up probability is 0.095. On the other hand, for $\theta_0 = 1/0.8$, 79.5% of the realizations produced a purely non-invertible model with an empirical pile-up probability of 0.109. Both objective functions do a reasonably good job of discriminating between invertible and non-invertible models, with a performance edge going to the exact likelihood. In terms of root mean square error, the performance of $\hat{\theta}_n^{(E)}$ is superior to $\hat{\theta}_n^{(J)}$ as θ_0 moves away from the unit circle.

Remark. The LAD estimate of θ_0 is obtained by minimizing the objective function given in (2.2) with $z_{init} = 0$. Although we have not considered the asymptotic pile-up in this case, the estimator does not perform as well as $\hat{\theta}_n^{(J)}$ and $\hat{\theta}_n^{(E)}$. For example, in simulation results, not reported here, the rmse of the LAD estimator tended to be twice as large as the rmse for the exact MLE.

References

[1] ANDERSON, T. W. AND TAKEMURA, A. (1986). Why do noninvertible estimated moving averages occur? *Journal of Time Series Analysis* **7** 235–254.

[2] BREIDT, F. J. AND DAVIS, R. A. (1992). Time-reversibility, identifiably, and independence of innovations for stationary time series. *Journal of Time Series Analysis* **13** 377–390.

[3] BROCKWELL, P. J. AND DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd Edition. Springer-Verlag, New York.

[4] Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16** 367–401.

[5] Chen, M., Davis, R. A. and Dunsmuir, W. T. M. (1995). Inference for MA(1) processes with a root on or near the unit circle. Invited paper in *Probability and Mathematical Statistics, Issue in Honour of Neyman's 100 Birthday* **15** 227–242.

[6] Davis, R. A. and Dunsmuir, W. T. M. (1996). Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. *Econometric Theory* **12** 1–29.

[7] Tanaka, K. (1996). *Time Series Analysis. Nonstationary and Noninvertible Distribution Theory.* Wiley, New York.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 20–47

DOI: 10.1214/074921706000000932

Recursive estimation of possibly misspecified MA(1) models: Convergence of a general algorithm

James L. Cantor[1] **and David F. Findley**[2]

Science Application International Corporation and U.S. Census Bureau

Abstract: We introduce a recursive algorithm of conveniently general form for estimating the coefficient of a moving average model of order one and obtain convergence results for both correct and misspecified MA(1) models. The algorithm encompasses Pseudolinear Regression (PLR—also referred to as AML and RML_1) and Recursive Maximum Likelihood (RML_2) without monitoring. Stimulated by the approach of Hannan (1980), our convergence results are obtained indirectly by showing that the recursive sequence can be approximated by a sequence satisfying a recursion of simpler (Robbins-Monro) form for which convergence results applicable to our situation have recently been obtained.

1. Introduction and overview

Our focus is on estimating the coefficient θ of an invertible scalar moving average model of order 1 (MA(1)),

$$y_t = \theta e_{t-1} + e_t \tag{1.1}$$

where e_t is treated as an unobserved, constant-variance martingale-difference process. We do not assume the series $y_t, -\infty < t < \infty$ from which the observations come is correctly modeled by (1.1). They can come from any invertible autoregressive moving average (ARMA) model or from more general models; see Section 2. What we seek is a θ that minimizes the loss function

$$\bar{L}(\theta) = E[(y_t - y_{t|t-1}(\theta))^2] = E[e_t^2(\theta)] \tag{1.2}$$

where $e_t(\theta) = y_t - y_{t|t-1}(\theta)$ and $y_{t|t-1}(\theta)$ is the one-step-ahead-prediction of y_t from $y_s, -\infty < s \le t-1$ based on the model defined by θ (see (2.7) below). We define *optimal* estimation procedures to be those whose sequence of estimates θ_t minimizes (1.2) in the limit. This is a property of (nonrecursive) maximum likelihood-type estimates of θ, see Pötscher [23].

In this article, we analyze a continuously indexed family of recursive procedures for estimating θ. Recursive procedures form an estimate θ_t for time t using the observation y_t at time t, the estimate θ_{t-1} for $t-1$ and other recursively defined quantities. Our family encompasses two standard algorithms, Recursive Maximum

[1]Science Applications International Corporation (SAIC), 4001 North Fairfax Drive, Suite 250, Arlington, VA 22203, e-mail: james.l.cantor@saic.com

[2]U.S. Census Bureau, Statistical Research Division, Room 3000-4, Washington, DC 20233-9100, e-mail: david.f.findley@census.gov

AMS 2000 subject classifications: primary 62M10; secondary 62L20.

Keywords and phrases: time series, Robbins-Monro, PLR, AML, RML_1, RML_2, misspecified models.

Likelihood (RML) which is referred to throughout as RML_2 [12, 21], and the simpler Pseudolinear Regression (PLR) [21]—also known as Approximate Maximum Likelihood (AML) [24] and RML_1 [11, 20]. More specifically, our general recursive algorithm generating θ_t depends on an index $\beta, 0 \le \beta \le 1$. The algorithm reduces to PLR when $\beta = 0$ and to RML_2 when $\beta = 1$.

Our main convergence result, Theorem 4.1, is obtained by constructing an approximating sequence $\hat{\theta}_t$ for which $\theta_t - \hat{\theta}_t \xrightarrow{a.s.} 0$ holds and which satisfies a Robbins-Monro recursion,

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \delta_t f(\hat{\theta}_{t-1}, \beta) + \delta_t \gamma_t \,, \tag{1.3}$$

in which $\gamma_t \xrightarrow{a.s.} 0$ and $\delta_t > 0, \delta_t \xrightarrow{a.s.} 0$, $\sum_{k=0}^{\infty} \delta_k = \infty$ a.s., and

$$f(\theta, \beta) = -\int_{-\pi}^{\pi} \frac{e^{i\omega} + \beta\theta}{|(1+\theta e^{i\omega})(1+\theta\beta e^{i\omega})|^2} g_y(\omega) d\omega. \tag{1.4}$$

Here $\xrightarrow{a.s.}$ denotes almost sure convergence (convergence with probability one) and $g_y(\omega)$ denotes the spectral density of the time series y_t. Note that when $\beta = 0$, then

$$f(\theta, 0) = -\int_{-\pi}^{\pi} \frac{e^{i\omega}}{|(1+\theta e^{i\omega})|^2} g_y(\omega) d\omega = -E[e_{t-1}(\theta) e_t(\theta)] \,, \tag{1.5}$$

and when $\beta = 1$, then

$$f(\theta, 1) = -\int_{-\pi}^{\pi} \frac{e^{i\omega} + \theta}{|(1+\theta e^{i\omega})^2|^2} g_y(\omega) d\omega = \frac{1}{2} \frac{d}{d\theta} E[e_t^2(\theta)] = \frac{1}{2} \bar{L}'(\theta) \tag{1.6}$$

where $\bar{L}'(\theta)$ denotes the first derivative of $\bar{L}(\theta)$. We then apply a result of Fradkov implicit in [8], as extended and corrected by Findley [9], to show that $\hat{\theta}_t$ converges to $\{\theta \in \Theta \colon f(\theta, \beta) = 0\}$ where Θ is the open interval (-1,1) of real θ with $|\theta| < 1$. (A similar result is implicit in proofs of Theorems 2.2.2–2.2.3 of Chen [7].) Hence, for $\beta = 0$, $\theta_t \xrightarrow{a.s.} \{\theta \in \Theta \colon E[e_{t-1}(\theta) e_t(\theta)] = 0\}$ and for $\beta = 1$, $\theta_t \xrightarrow{a.s.} \{\theta \in \Theta \colon \bar{L}'(\theta) = 0\}$. Here and below, θ_t convergence a.s. to a set means that except on a set of $\xi \in \Xi$ with probability zero, every cluster point of $\theta_t(\xi)$ is an element of the set.

In the incorrect model situation, in which $g_y(\omega)$ is not proportional to $|1+\theta e^{i\omega}|^2$, for examples we have analyzed [5], these zero sets will be disjoint, establishing that PLR converges to different values than RML_2. Consequently, under the assumptions of Theorem 4.1, we recover the results of Cantor [4] that were given in separate theorems and proofs, establishing that, for certain families of AR(1) and MA(2) processes, RML_2 estimates of θ in the model (1.1) converge to an optimal limit (a minimizer of (1.2)) whereas PLR estimates converge to a suboptimal limit [4, 5].

When the data come from an invertible MA(1) model, it is known that PLR and monitored versions of RML_2 can provide strongly consistent estimates of θ [4, 11, 17, 19]. More generally, in the correct model situation for ARMAX models, i.e., ARMA models with an exogenous input, Lai and Ying [17] provided a rigorous proof of strong consistency of PLR (under a positive real condition on the MA polynomial) and also of a monitored version of RML_2 whose monitoring scheme involves non-linear projections and an intermittently used recursive estimator for which consistency has already been established. In Section 4 of [19], Lai and Ying consider a simpler modification of RML_2 in which, for monitoring, only auxiliary consistent recursive estimates are used. They present detailed outlines of proofs of strong consistency and asymptotic normality of the estimates from this new

monitored RML_2 scheme. The construction of Section II of [18] can be used to obtain auxiliary recursive estimates with the properties required.

There is a rather comprehensive theory of recursive estimation of autoregressive (AR) models, encompassing certain incorrect model situations for algorithms like PLR (see e.g., [6]). There are, however, no published convergence results with rigorous proofs for MA models in the incorrect model situation. Ljung's seminal work on the convergence of recursive algorithms [20, 21] mentions the incorrect model situation but provides only suggestive results (further discussed in Section 5).

This article has five sections. In Section 2, the assumptions on the data and some consequences for the MA(1) model are given. In Section 3, the general recursive algorithm is presented. The Convergence Theorem is stated and proved in Section 4. Required preliminary technical results are given in Section 4.1 and the proof of the theorem is provided in Section 4.2. Finally, Section 5 concludes the article with a brief discussion.

2. Assumptions

The observations $y_t, t \geq 1$ are assumed to come from a mean zero, covariance stationary scalar series, $y_t, -\infty < t < \infty$ defined on the probability space $(\Xi, \mathcal{F}, P)$. We use the following additional assumptions on the process y_t:

(D1) y_1 is nonzero with probability one; i.e., $P\{y_1^2 > 0\} = 1$.
(D2) The series has a linear representation

$$y_t = \sum_{s=0}^{\infty} \kappa_s \epsilon_{t-s} \text{ such that } \kappa_0 = 1 \text{ and } \sum_{s=0}^{\infty} |\kappa_s| < \infty \tag{2.1}$$

in which $\kappa(z) = \sum_{s=0}^{\infty} \kappa_s z^s$ is nonzero for $|z| \leq 1$ and $\{\epsilon_t\}$ is a martingale-difference sequence (m.d.s.) with respect to the sequence of sigma fields $\mathcal{F}_t = \sigma(y_s, -\infty < s \leq t)$. Thus $E[\epsilon_t | \mathcal{F}_{t-1}] = 0$. By a result of Wiener [25, Theorem VI 5.2], $\kappa(z)^{-1} = \sum_{s=0}^{\infty} \beta_s z^s$ with $\sum_{s=0}^{\infty} |\beta_s| < \infty$, whence

$$\epsilon_t = \sum_{s=0}^{\infty} \beta_s y_{t-s} \ (\beta_0 = 1)\,. \tag{2.2}$$

(D3) The conditional variance $E[\epsilon_t^2 | \mathcal{F}_{t-1}]$ is constant almost surely; i.e., $E[\epsilon_t^2 | \mathcal{F}_{t-1}] = \sigma_\epsilon^2$ a.s. Equivalently, $E[\epsilon_t^2] = \sigma_\epsilon^2$ and $\epsilon_t^2 - \sigma_\epsilon^2$ is a m.d.s. with respect to the $\mathcal{F}_t$.
(D4) $\{\epsilon_t\}$ is bounded a.s.; $\sup_t |\epsilon_t| \leq K$ a.s. for some $K < \infty$.

From (D2)–(D3), the spectral density $g_y(\omega)$ can be expressed as

$$g_y(\omega) = \frac{\sigma_\epsilon^2}{2\pi} \left|\kappa(e^{i\omega})\right|^2 \text{ where } \kappa(e^{i\omega}) = \sum_{j=0}^{\infty} \kappa_j e^{ij\omega}\,, \tag{2.3}$$

and

$$0 < m \leq g_y(\omega) \leq M < \infty \text{ for all } -\pi \leq \omega \leq \pi \tag{2.4}$$

for positive constants m and M. The series y_t is an invertible ARMA process if and only if $\kappa(z)$ is a rational function.

Assumption (D4) is used extensively in the proof of the convergence theorem, Theorem 4.1, in Section 4.

Under (D2)–(D4), we can apply, for example, the First Moment Bound Theorem of Findley and Wei [10] to show that $t^{-1}\sum_{s=j+1}^{t}(y_s y_{s-j} - \gamma_j^y) \xrightarrow{a.s.} 0$. Hence, from the particular case $y_t = \epsilon_t$ in (2.1) and $j = 0$,

$$t^{-1}\sum_{s=1}^{t}\epsilon_s^2 \xrightarrow{a.s.} \sigma_\epsilon^2\,. \tag{2.5}$$

We consider models for y_t of the invertible, stationary first-order moving-average type (MA(1)) given by

$$y_t = \theta e_{t-1} + e_t,\ \ -\infty < t < \infty\,. \tag{2.6}$$

For a given coefficient θ such that $|\theta| < 1$, the difference equation (2.6) is satisfied with $e_t = e_t(\theta)$ given by the mean zero, covariance stationary one-step-ahead-prediction-error series,

$$e_t(\theta) = (1+\theta B)^{-1}y_t = \sum_{j=0}^{\infty}(-\theta)^j y_{t-j} = y_t - y_{t|t-1}(\theta)\,, \tag{2.7}$$

from the MA(1) predictor $y_{t|t-1}(\theta) = -\sum_{j=1}^{\infty}(-\theta)^j y_{t-j}$, see (5.1.21) of [3]. Here B is the backshift operator; i.e., $By_t = y_{t-1}$. The coefficient θ is referred to as the MA *coefficient*. Thus,

$$y_t = e_t(\theta) + \theta e_{t-1}(\theta)\,. \tag{2.8}$$

The infinite series in (2.7) converges in mean square and, from (D4) and the representation (2.1), also almost surely. Thus, $e_t(\theta)$ represents the optimal one-step-ahead-prediction-error process from the perspective of the model (2.6). The model (2.6) is correct if $e_t(\theta)$ coincides (a.s.) with the m.d.s. ϵ_t in (2.2), in which case $\beta_s = (-\theta)^s, k \geq 0$. Whether or not the model is correct for any θ, forecast errors $e_t(\theta)$ appearing in loss functions such as (1.2) and elsewhere are calculated as in (2.7). We emphasize that (2.1) allows data processes far more general than MA(1) processes. In particular, the z-transform, $\sum_{s=0}^{\infty}\kappa_s z^s$ is not required to be rational. For example, time series conforming to the exponential models of Bloomfield [2] have non-rational $\kappa(z)$ without zeroes in $|z| \leq 1$.

Let $\Theta = (-1, 1)$. From (2.7), the spectral density of $e_t(\theta)$ is $g_e(\theta, \omega) = g_y(\omega)\cdot|1+\theta e^{i\omega}|^{-2}$, so for $\bar{L}(\theta)$ defined by (1.2), we have

$$\bar{L}(\theta) = \int_{-\pi}^{\pi}\frac{g_y(\omega)}{|1+\theta e^{i\omega}|^2}d\omega\,. \tag{2.9}$$

By (2.4) and the continuity of $g_y(\omega)$, $\bar{L}(\theta)$ is positive, infinitely differentiable, and nonconstant on the interior of $[-1, 1]$, i.e., on Θ, and infinite at the endpoints. Therefore it has a minimum value over $[-1, 1]$ and

$$\Theta^* \equiv \left\{\theta \in [-1,1]\colon\ \theta = \arg\min_{\theta\in[-1,1]}\bar{L}(\theta)\right\}\,, \tag{2.10}$$

is a subset of $[-K, K]$ for some $0 < K < 1$. Also $\Theta^* \subseteq \Theta_0^* = \{\theta \in \Theta\colon \bar{L}'(\theta) = 0\}$. We are interested in a.s bounded random recursive sequences $\theta_t = \theta_t(\xi)$ that converge

a.s. to Θ^* or at least to Θ_0^*. If Θ_0^* contains only one point, θ_0^*, then θ_t converges to θ_0^* a.s. Our results will establish convergence of the sequence of estimates θ_t defined by the general algorithm presented below to the set of zeroes of $f(\theta, \beta)$ defined by (1.4).

3. The general recursive algorithm

For $0 \le \beta \le 1$, we define a general recursion for estimating the MA coefficient θ of (1.1):

$$\theta_t = \theta_{t-1} + \bar{P}_t^{-1} \frac{1}{t} \phi_{t-1} e_t; \;\; \theta_1 = 0, \, t \ge 2 \,, \tag{3.1a}$$

$$\bar{P}_t = \frac{1}{t} \sum_{s=1}^{t-1} \phi_s^2 = \bar{P}_{t-1} + \frac{1}{t} [\phi_{t-1}^2 - \bar{P}_{t-1}]; \;\; \bar{P}_1 = 0; \, t \ge 2 \,, \tag{3.1b}$$

$$e_t = y_t - \theta_{t-1} e_{t-1}; \;\; e_1 = y_1, \, t \ge 2 \,, \tag{3.1c}$$

$$\phi_t = x_t - \theta_{t-1} \phi_{t-1}; \;\; \phi_1 = x_1, \, t \ge 2 \,, \tag{3.1d}$$

$$x_t = y_t - \beta \theta_{t-1} x_{t-1}; \;\; x_1 = y_1, \, t \ge 2 \,. \tag{3.1e}$$

From (3.1a), it follows for $0 \le s \le t-1, t \ge 2$ that

$$\theta_{t-s} = \theta_t - \sum_{l=0}^{s-1} (t-l)^{-1} \bar{P}_{t-l}^{-1} \phi_{t-l-1} e_{t-l} \,, \tag{3.2}$$

where $\sum_{l=0}^{-1}(\cdot) \equiv 0$. From (3.1e),

$$x_t = \sum_{s=0}^{t-1} (-\beta)^s \left(\prod_{i=1}^{s} \theta_{t-i} \right) y_{t-s} \tag{3.3}$$

where $\prod_{i=1}^{0}(\cdot) \equiv 1$. Next, let $z_1 = e_1$ and, for $t \ge 2$,

$$z_t = e_t + \theta_{t-1} \phi_{t-1} \,. \tag{3.4}$$

The value of the parameterization with β is that it enables us to simultaneously obtain results for two important algorithms. When $\beta = 0$, then $x_t = y_t$ from which it follows that $\phi_t = e_t$ and $z_t = y_t$ and therefore (3.1a)–(3.1e) is PLR (AML, RML_1)[11, 20, 21, 24]. When $\beta = 1$, then $x_t = e_t$ and $\phi_t = e_t - \theta_{t-1}\phi_{t-1}$ and thus (3.1a)–(3.1e) is RML_2 [12, 21] without monitoring to ensure that each estimate θ_t is in $\Theta = (-1, 1)$.

For any β, these θ_t can be expressed in the form of a regression estimate:

$$\theta_t = \left\{ \sum_{s=2}^{t} \phi_{s-1}^2 \right\}^{-1} \sum_{s=2}^{t} z_s \phi_{s-1}, \;\; t \ge 2 \,. \tag{3.5}$$

An induction argument for (3.5) goes as follows. Set $P_t = t\bar{P}_t = \sum_{s=2}^{t} \phi_{s-1}^2$. Note that from (D1), $P_t > 0$ for all $t > 1$ and therefore P_t^{-1} exists a.s. From (3.1a)–(3.1e) and (3.4), $\theta_2 = \left(1/2\phi_1^2\right)^{-1} 1/2(z_2\phi_1)$, which is (3.5) for $t = 2$. Suppose then it is true for some $t \ge 2$; i.e.,

$$P_t \theta_t = \sum_{s=2}^{t} z_s \phi_{s-1} \,. \tag{3.6}$$

Then

$$\begin{aligned} P_{t+1}\theta_{t+1} &= P_{t+1}(\theta_t + P_{t+1}^{-1}\phi_t e_{t+1}) = (P_t + \phi_t^2)\theta_t + \phi_t e_{t+1} \\ &= \sum_{s=2}^{t} z_s\phi_{s-1} + \phi_t(\phi_t\theta_t + e_{t+1}) \text{ (from the induction hypothesis (3.6))} \\ &= \sum_{s=2}^{t} z_s\phi_{s-1} + \phi_t z_{t+1} = \sum_{s=2}^{t+1} z_s\phi_{s-1}\,. \end{aligned}$$

Hence, (3.5) is true for $t+1$ and by induction therefore for all t.

For use below, we define the stationary analogues $e_t(\theta), x_t(\theta), \phi_t(\theta)$ and $z_t(\theta)$ of e_t, x_t, ϕ_t and z_t:

$$e_t(\theta) = (1+\theta B)^{-1} y_t\,, \tag{3.7}$$

$$x_t(\theta) = (1+\theta\beta B)^{-1} y_t = \sum_{j=0}^{\infty} (-\beta\theta)^j y_{t-j}\,, \tag{3.8}$$

$$\begin{aligned} \phi_t(\theta) &= (1+\theta B)^{-1} x_t(\theta) = \sum_{j=0}^{\infty} (-\theta)^j x_{t-j}(\theta) \\ &= (1+\theta B)^{-1}(1+\theta\beta B)^{-1} y_t\,, \end{aligned} \tag{3.9}$$

so $\phi_t(\theta) = e_t(\theta)$ when $\beta = 0$. From (3.7)–(3.9),

$$z_t(\theta) = e_t(\theta) + \theta\phi_{t-1}(\theta) = [(1+\theta B)^{-1} + \theta B(1+\theta B)^{-1}(1+\theta\beta B)^{-1}] y_t\,. \tag{3.10}$$

From (3.7)–(3.10),

$$E[\phi_t^2(\theta)] = \int_{-\pi}^{\pi} \frac{1}{|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})|^2} g_y(\omega) d\omega\,, \tag{3.11}$$

$$\begin{aligned} E[\phi_{t-1}(\theta) e_t(\theta)] &= \int_{-\pi}^{\pi} \frac{e^{i\omega}}{(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})} \frac{1}{(1+\theta e^{-i\omega})} g_y(\omega) d\omega \\ &= \int_{-\pi}^{\pi} \frac{e^{i\omega} + \beta\theta}{|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})|^2} g_y(\omega) d\omega\,, \end{aligned} \tag{3.12}$$

and

$$E[z_t(\theta)\phi_{t-1}(\theta)] = \int_{-\pi}^{\pi} \frac{e^{i\omega} + \theta(1+\beta)}{|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})|^2} g_y(\omega) d\omega\,. \tag{3.13}$$

From (1.4) and (3.12), $E[\phi_{t-1}(\theta) e_t(\theta)] = -f(\theta, \beta)$. Let $e_t'(\theta) = de_t(\theta)/d\theta$. Then, from (3.7),

$$-e_t'(\theta) = \frac{B}{1+\theta B} e_t(\theta) = \frac{B}{(1+\theta B)^2} y_t\,. \tag{3.14}$$

Since

$$\frac{1}{2}\frac{d}{d\theta} E[e_t^2(\theta)] = E[e_t'(\theta) e_t(\theta)]\,,$$

from (2.9) and (3.14), the derivative of $\bar{L}(\theta)$, $\bar{L}'(\theta)$, is obtained from

$$\begin{aligned} -\frac{1}{2}\bar{L}'(\theta) &= E[-e_t'(\theta) e_t(\theta)] = \int_{-\pi}^{\pi} \frac{e^{i\omega}}{(1+\theta e^{i\omega})^2} \frac{1}{(1+\theta e^{-i\omega})} g_y(\omega) d\omega \\ &= \int_{-\pi}^{\pi} \frac{e^{i\omega} + \theta}{|(1+\theta e^{i\omega})^2|^2} g_y(\omega) d\omega\,, \end{aligned} \tag{3.15}$$

which is (3.12) with $\beta = 1$, verifying (1.6).

As a consequence of (2.4), we note that since $|z| \le K^* < 1$ implies $0 < 1 - K^* \le |1 - z| \le 1 + K^*$, for (3.11) with $|\theta| \le K^* < 1$ we have

$$\frac{m}{(1+K^*)^4} \le \int_{-\pi}^{\pi} \frac{1}{|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})|^2} g_y(\omega)d\omega \le \frac{M}{(1-K^*)^4}. \tag{3.16}$$

4. The convergence theorem

The following result is a generalization of the PLR and RML_2 results proved in [4] for MA(1) models.

Theorem 4.1 (Convergence theorem). *Consider a series y_t for which (D1)–(D4) hold. For each β such that $0 \le \beta \le 1$, assume that the recursive sequence defined by (3.1a)–(3.1e) is such that, for some random $k^* = k^*(\xi)$ and $K^* = K^*(\xi)(\xi \in \Xi)$ satisfying $0 \le k^* < \infty$ and $0 < K^* < 1$, it holds almost surely that $|\theta_{t+k^*}| \le K^*$ for all t. Then for $f(\theta, \beta)$ as in (1.4):*

(a) *The sequence $\hat{\theta}_t$ defined for $t \ge 1$ by*

$$\begin{aligned}\hat{\theta}_t = &\left[\frac{1}{t}\sum_{s=1}^{t}\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{s+k^*}e^{i\omega})(1+\beta\theta_{s+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right]^{-1} \\ &\times \frac{1}{t}\sum_{s=1}^{t}\int_{-\pi}^{\pi} \frac{\cos\omega + (1+\beta)\theta_{s+k^*}}{|(1+\theta_{s+k^*}e^{i\omega})(1+\beta\theta_{s+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\end{aligned} \tag{4.1}$$

has the property that $\theta_t - \hat{\theta}_t \xrightarrow{a.s.} 0$. Hence, with probability one, there is a $t_0(\xi) \ge 1$ such that $|\hat{\theta}_t| \le (1+K^*)/2 < 1$ holds for all $t \ge t_0(\xi)$.

(b) *For all $t > t_0(\xi)$, $\hat{\theta}_t$ satisfies a Robbins-Monro recursion,*

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \delta_t f(\hat{\theta}_{t-1}, \beta) + \delta_t \gamma_t\,, \tag{4.2}$$

with $\gamma_t \xrightarrow{a.s.} 0$, $\delta_t > 0$ a.s., $\delta_t \xrightarrow{a.s.} 0$, and $\sum_{s=t_0+1}^{\infty} \delta_s = \infty$ a.s. where $f(\theta, \beta)$ has the formula (1.4).

(c) *From* (a) *and* (b), *it follows that, with $\Theta = (-1, 1)$, the sequence θ_t converges a.s. to the compact set*

$$\Theta_0^\beta = \{\theta \in \Theta : f\,(\theta, \beta) = 0\} \tag{4.3}$$

in the sense that, on a probability one event Ξ_0 that does not depend on β, for each $\xi \in \Xi_0$, the cluster points of $\theta_t(\xi)$ are contained in Θ_0^β. Further, when y_t is an invertible ARMA process, then Θ_0^β is finite, and $\theta(\xi) = \lim_{t\to\infty} \theta_t(\xi)$ exists for every $\xi \in \Xi_0$.

Note from (3.5), (3.11) and (3.13) that the assertion $\theta_t - \hat{\theta}_t \xrightarrow{a.s.} 0$ in part (a) of Theorem 4.1 can be formulated as the assertion that

$$\begin{aligned}&\left\{\frac{1}{t}\sum_{s=1}^{t}\phi_{s-1}^2\right\}^{-1} \frac{1}{t}\sum_{s=1}^{t} z_s\phi_{s-1} \\ &\quad - \left[\frac{1}{t}\sum_{s=1}^{t} E[\phi_t^2(\theta_{s+k^*})]\right]^{-1} \frac{1}{t}\sum_{s=1}^{t} E[z_t(\theta_{s+k^*})\phi_{t-1}(\theta_{s+k^*})]\end{aligned}$$

tends to zero a.s. In the expression above, $\phi_0 = 0$ and expectation is taken before evaluation at θ_{s+k^*}.

The proof of Theorem 4.1, given in Section 4.2. In [5], we provide complete results concerning the existence of k^* and K^* with the required properties for several incorrect model examples as well as for the correct model situation for $\beta = 0$ (PLR) and provide more limited results for the case $\beta = 1$ (RML_2) with a particular monitoring scheme. For the latter case, we also report on simulation results which demonstrate the existence of the variates k^*, K^* as in Theorem 4.1 with the consequence that monitoring becomes unnecessary for sufficiently large t. In the correct model case $y_t = \theta\epsilon_{t-1} + \epsilon_t$ with i.i.d. ϵ_t, Lai and Ying [19] show for their monitored RML_2 that this happens a.s. and the conclusions of Theorem 4.1 concerning our approximating sequence (4.1) apply.

4.1. Preliminary results

Here we present some needed technical results. We first quote, without proof, a powerful result from martingale theory [17, Lemma 1, part (i)]. Unless specified otherwise, all limits (liminfs, limsups, etc.) are with respect to t and for simplicity the $t \to \infty$ will be usually suppressed.

Proposition 4.1. *Let $\{\tilde\epsilon_t\}$ be a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{F}_t\}$ such that $\sup_t E[|\tilde\epsilon_t|^{2p}|\mathcal{F}_{t-1}] < \infty$ holds a.s. for some $p > 1$. Let $\tilde z_t$ be an $\mathcal{F}_{t-1}$-measurable random variable for every t. Then $\sum_{s=1}^t \tilde z_s\tilde\epsilon_s$ converges almost surely on $\{\sum_{s=1}^\infty \tilde z_s^2 < \infty\}$, and for every $\eta > 1/2$,*

$$\frac{\left(\sum_{s=1}^t \tilde z_s\tilde\epsilon_s\right)}{\left(\sum_{s=1}^t \tilde z_s^2\right)^\eta} \xrightarrow{a.s.} 0 \ \text{ on } \ \left\{\sum_{s=1}^\infty \tilde z_s^2 = \infty\right\}.$$

Since

$$\frac{1}{t}\sum_{s=1}^t \tilde z_s\tilde\epsilon_s = \left\{\frac{\sum_{s=1}^t \tilde z_s\tilde\epsilon_s}{\sum_{s=1}^t \tilde z_s^2}\right\}\frac{1}{t}\sum_{s=1}^t \tilde z_s^2\,,$$

it is clear that a corollary of this Proposition is

Proposition 4.2. *Under the assumptions of Proposition 4.1, if $\limsup t^{-1} \times \sum_{s=1}^t \tilde z_s^2 < \infty$ a.s., then $t^{-1}\sum_{s=1}^t \tilde z_s\tilde\epsilon_s \xrightarrow{a.s.} 0$.*

Recall from (2.1) that $y_t = \epsilon_t + \sum_{s=1}^\infty \kappa_s\epsilon_{t-s}$ since $\kappa_0 = 1$. A second consequence of Proposition 4.1 is

Proposition 4.3. *Suppose that the m.d.s. ϵ_t in (D2) is such that $\sup_t E[|\epsilon_t|^{2p}|\mathcal{F}_{t-1}] < \infty$ holds a.s. for some $p > 1$. Then for any sequence $\hat y_t = y_t - \tilde y_{t-1}$ in which $\tilde y_{t-1}$ is $\mathcal{F}_{t-1}$-measurable, it holds that $\liminf t^{-1}\sum_{s=1}^t \hat y_s^2 \ge \sigma_\epsilon^2$ a.s., where $\sigma_\epsilon^2 = E[\epsilon_t^2]$.*

Proof. From (2.1), $\hat y_t = y_t - \tilde y_{t-1} = \epsilon_t + \tilde z_t$ where $\tilde z_t = -\tilde y_{t-1} + \sum_{s=1}^\infty \kappa_s\epsilon_{t-s}$ is $\mathcal{F}_{t-1}$-measurable since $\sum_{s=1}^\infty \kappa_s\epsilon_{t-s}$ is $\mathcal{F}_{t-1}$-measurable by (2.2) and $\tilde y_{t-1}$ is $\mathcal{F}_{t-1}$-measurable by assumption. Then

$$\begin{aligned}\frac{1}{t}\sum_{s=1}^t \hat y_s^2 &= \frac{1}{t}\sum_{s=1}^t \epsilon_s^2 + \frac{2}{t}\sum_{s=1}^t \epsilon_s\tilde z_s + \frac{1}{t}\sum_{s=1}^t \tilde z_s^2 \\ &= \frac{1}{t}\sum_{s=1}^t \epsilon_s^2 + \left\{2\frac{\sum_{s=1}^t \epsilon_s\tilde z_s}{\sum_{s=1}^t \tilde z_s^2} + 1\right\}\frac{1}{t}\sum_{s=1}^t \tilde z_s^2\,.\end{aligned} \tag{4.4}$$

Consider first the event that $\sum_{s=1}^{t}\tilde{z}_s^2 \xrightarrow{a.s.} l < \infty$. Then $t^{-1}\sum_{s=1}^{t}\tilde{z}_s^2 \xrightarrow{a.s.} 0$ and, by the preceding Proposition, $t^{-1}\sum_{s=1}^{t}\epsilon_s\tilde{z}_s \xrightarrow{a.s.} 0$. Hence, from (2.5) and the first equation in (4.4), $\lim t^{-1}\sum_{s=1}^{t}\hat{y}_s^2 = t^{-1}\sum_{s=1}^{t}\epsilon_s^2 = \sigma_\epsilon^2$ so the assertion holds in this event. In the complementary event, $\sum_{s=1}^{t}\tilde{z}_s^2 \xrightarrow{a.s.} \infty$, from (4.4), it follows that

$$\begin{aligned}\liminf \frac{1}{t}\sum_{s=1}^{t}\hat{y}_s^2 &= \liminf\left(\frac{1}{t}\sum_{s=1}^{t}\epsilon_s^2 + \left\{2\frac{\sum_{s=1}^{t}\epsilon_s\tilde{z}_s}{\sum_{s=1}^{t}\tilde{z}_s^2}+1\right\}\frac{1}{t}\sum_{s=1}^{t}\tilde{z}_s^2\right)\\ &= \sigma_\epsilon^2 + \liminf\left(\left\{2\frac{\sum_{s=1}^{t}\epsilon_s\tilde{z}_s}{\sum_{s=1}^{t}\tilde{z}_s^2}+1\right\}\frac{1}{t}\sum_{s=1}^{t}\tilde{z}_s^2\right)\ a.s.\end{aligned} \tag{4.5}$$

By Proposition 4.1, $\sum_{s=1}^{t}\epsilon_s\tilde{z}_s/\sum_{s=1}^{t}\tilde{z}_s^2 \xrightarrow{a.s.} 0$. Hence, the second expression in (4.5) is nonnegative, and the proof is complete. □

Proposition 4.4. *Under (2.4), for each $\beta \in [0,1]$, the function $f(\theta,\beta)$ defined by (1.4) is infinitely differentiable on $\Theta = (-1,1)$, and Θ_0^β defined by (4.3) is a nonempty compact subset of Θ. In the case $\beta = 1$, Θ_0^1 contains the (nonempty) set of minimizers over Θ of $\bar{L}(\theta)$ defined by (2.9).*

Proof. The differentiability assertion follows from (2.4) via the dominated convergence theorem. Except for compactness of Θ_0^1, which will be discussed below, the assertions concerning $\bar{L}(\theta)$ and $f(\theta,1)$ were obtained subsequent to (2.10). The remaining assertions follow from the continuity of $f(\theta,\beta)$ and the limit properties

$$\lim_{\theta\to-1} f(\theta,\beta) = -\infty \tag{4.6}$$

and

$$\lim_{\theta\to 1} f(\theta,\beta) = \infty. \tag{4.7}$$

Indeed, from (4.6)–(4.7), for any $K > 0$ there exists an $0 < \epsilon(K,\beta) < 1$ such that $f(\theta,\beta) \le -K$ for all $\theta \in (-1,-1+\epsilon)$ and $f(\theta,\beta) \ge K$ for all $\theta \in (1-\epsilon,1)$. Therefore $f(\theta,\beta)$ must change sign over $[-1+\varepsilon, 1-\varepsilon]$. Hence $f(\theta,\beta)$ is non-constant and has a zero in this interval and, moreover, $\Theta_0^\beta \subseteq [-1+\varepsilon, 1-\varepsilon]$. Finally, since $f(\theta,\beta)$ is continuous on this interval, Θ_0^β is compact. An analogous argument applies to Θ_0^1.

To verify (4.6), we note that $g_y(\omega) = g_y(-\omega), -\pi \le \omega \le \pi$ yields

$$f(\theta,\beta) = -\int_{-\pi}^{\pi}\frac{\cos\omega+\beta\theta}{\left|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})\right|^2}g_y(\omega)\,d\omega.$$

Because $0 \le \beta < 1$, for $0 < \varepsilon < 1-\beta$ there is a $\delta = \delta(\varepsilon) \in (0,\pi)$ such that $\cos\omega + \beta\theta \ge \varepsilon$ whenever $|\omega| \le \delta$ and $-1 \le \theta \le 0$. For such ε, δ, we obtain

$$\begin{aligned}&\lim_{\theta\to-1}\int_{-\pi}^{\pi}\frac{\cos\omega+\beta\theta}{\left|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})\right|^2}g_y(\omega)\,d\omega\\ &\quad= \left\{\int_{-\pi}^{-\delta}+\int_{\delta}^{\pi}\right\}\frac{\cos\omega+\beta\theta}{\left|(1-e^{i\omega})(1-\beta e^{i\omega})\right|^2}g_y(\omega)\,d\omega \qquad (4.8)\\ &\qquad+\lim_{\theta\to-1}\int_{-\delta}^{\delta}\frac{\cos\omega+\beta\theta}{\left|(1+\theta e^{i\omega})(1+\beta\theta e^{i\omega})\right|^2}g_y(\omega)\,d\omega \qquad (4.9)\\ &\quad= \infty,\end{aligned}$$

because (4.8) is finite, whereas for (4.9) we have

$$\lim_{\theta\to-1}\int_{-\delta}^{\delta}\frac{\cos\omega+\beta\theta}{\left|\left(1+\theta e^{i\omega}\right)\left(1+\beta\theta e^{i\omega}\right)\right|^2}g_y\left(\omega\right)d\omega$$
$$\geq\varepsilon\, m\lim_{\theta\to-1}\int_{-\delta}^{\delta}\left|\left(1+\theta e^{i\omega}\right)\left(1+\beta\theta e^{i\omega}\right)\right|^{-2}d\omega=\infty.$$

This yields (4.6), and (4.7) follows by an analogous argument. □

Proposition 4.5. *Let y_t be an invertible ARMA process, then for each $\beta\in[0,1]$, the set $\Theta_0^{\beta}=\{\theta\in(-1,1):f(\theta,\beta)=0\}$ is finite.*

Proof. $\kappa\left(z\right)$ in (D2) has the form $\kappa\left(z\right)=\eta\left(z\right)/\phi\left(z\right)$ where $\eta\left(z\right)$ and $\phi\left(z\right)$ are polynomials, of degrees d_η and d_ϕ, respectively, having no common zeroes and having all zeros in $\{|z|>1\}$. Setting $z=e^{i\omega}$ and $h\left(z\right)=\left(1+\theta z\right)\left(1+\beta\theta z\right)$, we obtain from $dz=izd\omega$ that

$$\begin{aligned}-f(\theta,\beta)&=\int_{-\pi}^{\pi}\frac{e^{i\omega}+\beta\theta}{\left|\left(1+\theta e^{i\omega}\right)\left(1+\beta\theta e^{i\omega}\right)\right|^2}g_y\left(\omega\right)d\omega\\&=\frac{\sigma_\varepsilon^2}{2\pi i}\int_{|z|=1}\frac{\left(z+\beta\theta\right)\eta\left(z\right)\eta\left(z^{-1}\right)}{zh\left(z\right)h\left(z^{-1}\right)\phi\left(z\right)\phi\left(z^{-1}\right)}dz\\&=\frac{\sigma_\varepsilon^2}{2\pi i}\int_{|z|=1}z^{1+d_\phi-d_\eta}\frac{\left(z+\beta\theta\right)\eta\left(z\right)\left\{z^{d_\eta}\eta\left(z^{-1}\right)\right\}}{h\left(z\right)\left\{z^2h\left(z^{-1}\right)\right\}\phi\left(z\right)\left\{z^{d_\phi}\phi\left(z^{-1}\right)\right\}}dz.\end{aligned}$$

The function

$$w\left(z\right)=\sigma_\varepsilon^2z^{1+d_\phi-d_\eta}\frac{\left(z+\beta\theta\right)\eta\left(z\right)\left\{z^{d_\eta}\eta\left(z^{-1}\right)\right\}}{h\left(z\right)\left\{z^2h\left(z^{-1}\right)\right\}\phi\left(z\right)\left\{z^{d_\phi}\phi\left(z^{-1}\right)\right\}}$$

is nonzero on $\{|z|=1\}$ and has poles interior to the unit circle at $-\theta$, $-\beta\theta$, at the zeroes of $z^{d_\eta}\phi\left(z^{-1}\right)$, and, if $1+d_\phi-d_\eta<0$, also at 0. If $z_j, j=1,\ldots,n$ are the distinct poles in $\{z\colon|z|<1\}$, then, by the Residue Theorem of complex analysis, e.g., (4.7-10) of Henrici [13], it follows that

$$f\left(\theta,\beta\right)=-\sum_{j=1}^{n}\mathrm{Res}_{z=z_j}w\left(z\right),$$

where, if z_j is a pole of order $J\geq1$,

$$\mathrm{Res}_{z=z_j}w\left(z\right)=\frac{1}{(J-1)!}\lim_{z\to z_j}\frac{d^{J-1}}{dz^{J-1}}\left\{\left(z-z_j\right)^Jw\left(z\right)\right\}.$$

Thus each $\mathrm{Res}_{z=z_j}w\left(z\right)$ is a rational function of θ, and therefore the same is true of $f\left(\theta,\beta\right)$. Consequently, $f\left(\theta,\beta\right)=0$ holds for only finitely many θ in $(-1,1)$. □

The final preliminary result addresses convergence of a Robbins-Monro type recursion that will be applied to demonstrate convergence of the general recursive algorithm. It is a special case of a correction and extension by Findley [9] of a result that is implicit in the proof of a theorem of Fradkov presented in Derevitzkiĭ and Fradkov [8] for the case of monotonically decreasing δ_t. The result below is also implicit in the proofs of Theorem 2.2.2 and Corollary 2.2.1 of Chen [7] which cover the case of vector θ more completely than Findley [9].

Proposition 4.6. *Let $\hat{\theta}_t$, $t \geq t_0$ be a non-stochastic, real-valued sequence satisfying*

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \delta_t f(\hat{\theta}_{t-1}) + \delta_t \gamma_t, \ t > t_0$$

for some real-valued function $f(\theta)$, with γ_t, $t > t_0$ satisfying $\gamma_t \to 0$ and with δ_t, $t \geq t_0$ satisfying $\delta_t \geq 0$, $\delta_t \to 0$, and $\sum_{t=t_0+1}^{\infty} \delta_t = \infty$. Suppose there is a bounded open set $\tilde{\Theta}$ on which $f(\theta)$ is continuously differentiable and which is such that the sequence $\hat{\theta}_t$ enters $\tilde{\Theta}$ infinitely often and has no cluster point on the boundary of $\tilde{\Theta}$. Then $\hat{\theta}_t$ is bounded, and its cluster points belong to $\tilde{\Theta}_0 = \{\theta \in \tilde{\Theta} \colon f(\theta) = 0\}$, i.e., $\hat{\theta}_t \to \tilde{\Theta}_0$. The set of cluster points is compact. If $\tilde{\Theta}_0$ is finite, then $\hat{\theta}_t$ converges to some $\theta \in \tilde{\Theta}_0$.

4.2. Proof of the convergence theorem

The proof of Theorem 4.1 follows from a set of technical lemmas and propositions given below. Proposition 4.7 provides a set of technical results needed to prove the Theorem's two main assertions: (i) the asymptotic equivalence of θ_t and the sequence $\hat{\theta}_t$ (Proposition 4.8) and (ii) (Proposition 4.9) the fact that $\hat{\theta}_t$ satisfies a.s. a Robbins-Monro recursion of the form considered in Proposition 4.6.

Hereafter, K or sometimes k (or these letters with decorations) will denote a generic upper bound (not always the same one) that is finite, or when it is random, finite a.s. A random K will be shown as $K(\xi)$ with $\xi \in \Xi$ on first appearance whenever the randomness is not immediately clear from context. Again, unless specified otherwise, all limits (liminfs, limsups, etc.) are with respect to t and usually the $t \to \infty$ will be omitted. The notation $o_{a.s.}(1)$ denotes convergence to zero with probability one.

Proposition 4.7. *Under the assumptions of Theorem 4.1, for the general recursive algorithm, the assertions* (a)–(c) *below follow:*

(a) $\liminf t^{-1} \sum_{s=1}^{t} \phi_s^2 \geq \sigma_\epsilon^2$ *a.s. and* $(t^{-1} \sum_{s=1}^{t} \phi_s^2)^{-1} \leq K(\xi) < \infty$, *and thus, from (3.1b),* $\bar{P}_t^{-1}$ *is bounded a.s.*

(b) *For* $t \geq 1$, $e_t = \sum_{j=0}^{\infty} \kappa_j^e(t)\epsilon_{t-j}; \phi_t = \sum_{j=0}^{\infty} \kappa_j^\phi(t)\epsilon_{t-j}; x_t = \sum_{j=0}^{\infty} \kappa_j^x(t)\epsilon_{t-j}$; *and* $z_t = \sum_{j=0}^{\infty} \kappa_j^z(t)\epsilon_{t-j}$ *where for every* j, $\kappa_j^e(t), \kappa_j^\phi(t), \kappa_j^x(t)$ *and* $\kappa_j^z(t)$ *are* $\mathcal{F}_{t-1}$*-measurable. Moreover, there exist* $\tilde{\kappa}_j$ *such that*

$$\max_j \{|\kappa_j^e(t)|, |\kappa_j^\phi(t)|, |\kappa_j^x(t)|, |\kappa_j^z(t)|\} \leq \tilde{\kappa}_j$$

and $\sum_j^{\infty} \tilde{\kappa}_j < \infty$ *a.s. Hence, the sequences* e_t, ϕ_t, x_t *and* z_t *are uniformly bounded a.s.*

(c) $\theta_t - \theta_{t-1} = o_{a.s.}(1)$.

Proof of (a). From (3.1d), $\phi_t = x_t - \theta_{t-1} e_{t-1} = y_t - \theta_{t-1}(\beta x_{t-1} + e_{t-1})$. Since $\theta_{t-1}(\beta x_{t-1} + e_{t-1})$ is $\mathcal{F}_{t-1}$-measurable, by Proposition 4.3,

$$\liminf t^{-1} \sum_{s=1}^{t} \phi_s^2 \geq \sigma_\epsilon^2 \ \ a.s. \tag{4.10}$$

Continuing, from (4.10), for any $0 < L_1 < \sigma_\epsilon^2$, there exists $t_0 = t_0(L_1, \xi)$ such that $t^{-1} \sum_{s=1}^{t} \phi_s^2 > L_1$ a.s. for all $t \geq t_0$. Let $L_2(\xi) \equiv \min_{1 \leq t < t_0} t^{-1} \sum_{s=1}^{t} \phi_s^2$. Then

$0 < L_2 < \infty$ a.s. This follows since t_0 is finite and ϕ_t is a finite valued sequence with probability one, hence $L_2 < \infty$. Moreover, since $\phi_1 = y_1$, under (D1) it follows that $L_2 > 0$ a.s. Hence, $(t^{-1}\sum_{s=1}^{t}\phi_s^2)^{-1} \leq \max\{L_1^{-1}, L_2^{-1}\} < \infty$ a.s. and the proof of part (a) is complete. □

Proof of (b). Set $\theta_0 = 0$. From $e_1 = y_1$ and $e_t = y_t - \theta_{t-1}e_{t-1}, t \geq 2$, it follows that $\kappa_j^e(1) = \kappa_j$ for all j, that $\kappa_0^e(t) = \kappa_0$ for all $t \geq 1$, and that $\kappa_j^e(t) = \kappa_j(t) - \theta_{t-1}\kappa_j^e(t-1)$ for all $t \geq 2, j \geq 1$. It follows by induction that

$$\kappa_j^e(t) = \sum_{l=0}^{\min(j,t-1)} (-1)^l \kappa_{j-l} \prod_{i=1}^{l} \theta_{t-i} \quad \text{where } \textstyle\prod_{i=1}^{0}(\cdot) \equiv 1\,. \tag{4.11}$$

Since for some k^* finite, $|\theta_{t+k^*}| < 1$ for all $t \geq 1$, we have that $|\theta_t| \leq K(\xi) < \infty$. First suppose that $K < 1$. Then from (4.11),

$$|\kappa_j^e(t)| \leq \sum_{l=0}^{\min(j,t-1)} |\kappa_{j-l}| \prod_{i=1}^{l} |\theta_{t-i}| \leq \sum_{l=0}^{j} K^l |\kappa_{j-l}|$$

and since $K < 1$, $\sum_{j=0}^{\infty}|\kappa_j^e(t)| \leq \sum_{j=0}^{\infty}\sum_{l=0}^{j} K^l|\kappa_{j-l}| = \sum_{l=0}^{\infty} K^l \sum_{p=0}^{\infty}|\kappa_p| < \infty$ where $p = j - l$. So the result holds for the case of $0 < K < 1$.

Otherwise, suppose $1 \leq K < \infty$. For all $t \geq k^*$, we have that $|\theta_t| \leq K^*(\xi) < 1$, so $K(\xi) = \lambda(\xi)K^*(\xi)$ for $\lambda > 1$. For simplicity of notation, replace K^* by ρ. We next show that $\prod_{i=1}^{l}|\theta_{t-i}| \leq \lambda^{k^*}\rho^l$ for $l \leq t$. First suppose $t \leq k^*$. Then $\prod_{i=1}^{l}|\theta_{t-i}| \leq \lambda^l\rho^l \leq \lambda^{k^*}\rho^l$. Next suppose $t > k^*$ and $l \leq t - k^*$. Then, $\prod_{i=1}^{l}|\theta_{t-i}| \leq \rho^l < \rho^l\lambda^{k^*}$ since $|\theta_{t-i}| \leq \rho$ for $1 \leq i \leq t - s^*$. Finally, suppose $t > k^*$ and $l > t - s^*$. Then since $l \leq t$,

$$\begin{aligned}\prod_{i=1}^{l}|\theta_{t-i}| &= \prod_{i=1}^{t-s^*}|\theta_{t-i}| \prod_{i=t-s^*+1}^{l}|\theta_{t-i}| \leq \rho^{t-s^*}\lambda^{l-(t-s^*)}\rho^{l-(t-s^*)} \\ &= \rho^l\lambda^{l-(t-s^*)} = \lambda^{k^*}\lambda^{l-t}\rho^l \leq \lambda^{k^*}\rho^l\,.\end{aligned}$$

Hence, generally $\prod_{i=1}^{l}|\theta_{t-i}| \leq \lambda^{k^*}\rho^l$. Setting $\kappa_j^e(\xi) = \lambda^{k^*}\sum_{l=0}^{j}\rho^l|\kappa_{j-l}|$, we have

$$|\kappa_j^e(t)| \leq \sum_{l=0}^{j}|\kappa_{j-l}| \prod_{i=1}^{l}|\theta_{t-i}| \leq \lambda^{k^*}\sum_{l=0}^{j}\rho^l|\kappa_{j-l}| = \kappa_j^e\,,$$

and since $|\rho| < 1$, $\sum_{j=0}^{\infty}\kappa_j^e < \infty$ a.s.

Next, from (3.3)

$$\kappa_j^x(t) = \sum_{l=0}^{\min(j,t-1)} (-\beta)^l \kappa_{j-l} \prod_{i=1}^{l}\theta_{t-i}\,, \tag{4.12}$$

and since $0 \leq \beta \leq 1$, an argument like that for e_t can be applied and to obtain the existence of a κ_j^x such that

$$|\kappa_j^x(t)| \leq \kappa_j^x \text{ and } \sum_{j=0}^{\infty}\kappa_j^x < \infty \, a.s. \tag{4.13}$$

Continuing, since $\phi_1 = x_1$ and $\phi_t = x_t - \theta_{t-1}\phi_{t-1}$ for $t \geq 2$, it follows similarly that

$$\kappa_j^{\phi}(t) = \sum_{l=0}^{\min(j,t-1)} (-1)^l \kappa_{j-l}^{x}(t) \prod_{i=1}^{l} \theta_{t-i} \,. \tag{4.14}$$

From (4.12) and (4.13), substituting $\kappa_j^x(t)$ for κ_j, the same kind of argument can be applied to (4.14) to yield

$$|\kappa_j^{\phi}(t)| \leq \kappa_j^{\phi} \text{ with } \textstyle\sum_{j=0}^{\infty} \kappa_j^{\phi} < \infty \, a.s. \tag{4.15}$$

Finally, for $t \geq 2$, we have, from $z_t = e_t + \theta_{t-1}\phi_{t-1}$,

$$\sum_{j=0}^{\infty} \kappa_j^z(t)\epsilon_{t-j} = \sum_{j=0}^{\infty} \kappa_j^e(t)\epsilon_{t-j} + \theta_{t-1} \sum_{j=0}^{\infty} \kappa_j^{\phi}(t-1)\epsilon_{t-1-j} \,,$$

for $t \geq 2$ from which it follows that

$$\kappa_j^z(t) = \kappa_j^e(t) + \theta_{t-1}\kappa_{j-1}^{\phi}(t-1) \,, \tag{4.16}$$

where $\kappa_{-1}^{\phi}(t) \equiv 0$. Since $\sup_t |\theta_t| < \infty$ a.s.,

$$|\kappa_j^z(t)| \leq \kappa_j^e + \sup_t |\theta_t| \kappa_{j-1}^{\phi} \;\; a.s.,$$

where $\kappa_{-1}^{\phi} \equiv 0$, so there is a κ_j^z such that $|\kappa_j^z(t)| \leq \kappa_j^z$ and $\sum_{j=0}^{\infty} |\kappa_j^z| < \infty$ a.s. for $t \geq 2$. Since $z_1 = e_1$, it thus follows that $\tilde{\kappa}_j = \max_j\{|\kappa_j^e|, |\kappa_j^{\phi}|, |\kappa_j^x|, |\kappa_j^z|\}$ satisfies $\sum_j^{\infty} \tilde{\kappa}_j < \infty$ a.s.

From this, we see that e_t, ϕ_t, x_t and z_t are bounded a.s. For example,

$$|\phi_t| = \left| \sum_{j=0}^{\infty} \kappa_j^{\phi}(t)\epsilon_{t-j} \right| \leq \sup_{-\infty<t<\infty} |\epsilon_t| \sum_{j=0}^{\infty} \tilde{\kappa}_j < \infty \;\; a.s.$$

From (4.11)–(4.12), (4.14) and (4.16), $\kappa_j^e(t), \kappa_j^{\phi}(t), \kappa_j^x(t)$ and $\kappa_j^z(t)$ are each $\mathcal{F}_{t-1}$-measurable for every j. Hence, part (b) of the Proposition is proved. □

Proof of (c). By parts (a) and (b), $|\theta_t - \theta_{t-1}| \leq t^{-1}\bar{P}_t^{-1}|e_t||\phi_{t-1}| \leq t^{-1}K(\xi)$ where $K(\xi) < \infty$ and thus part (c) follows and the proof of Proposition 4.7 is complete. □

Lemma 4.1. *Under the assumptions of Theorem 4.1, we have:*

(a) *If $\tilde{\kappa}_j(t)$ are $\mathcal{F}_{t-1}$-measurable such that $|\tilde{\kappa}_j(t)| \leq \tilde{\kappa}_j$ for $j \geq 0$, with $\sum_{j=0}^{\infty} \tilde{\kappa}_j < \infty$ a.s., then for all $p \geq 1$ and each $0 \leq j < \infty$,*

$$\frac{1}{t} \sum_{s=2}^{t} \left(\sum_{l=1}^{\min(j,s-1)} \tilde{\kappa}_{j-l}(s) \prod_{i=1}^{l} (s-i)^{-1} \bar{P}_{s-i}^{-1} \phi_{s-i-1} e_{s-i} \right)^p \xrightarrow{a.s.} 0 \,, \tag{4.17}$$

and

$$\frac{1}{t} \sum_{s=2}^{t} \left(\sum_{l=1}^{\min(j,s-1)} \tilde{\kappa}_{j-l}(s) \sum_{i=0}^{l} (s-i)^{-1} \bar{P}_{s-i}^{-1} \phi_{s-i-1} e_{s-i} \right)^p \xrightarrow{a.s.} 0 \,. \tag{4.18}$$

In particular,

$$\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{l=1}^{\min(j,s-1)}\tilde{\kappa}_{j-l}(s)\prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}\right)^{p}\epsilon_{s-j}^{2}\xrightarrow{a.s.}0\,. \tag{4.19}$$

(b) *For any* $0\le j<\infty$ *and* $i\le j$,

$$\frac{1}{t}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2\epsilon_{s-j}^2=\frac{1}{t}\sum_{s=i+1}^{t}(\kappa_j^{\phi}(s-i))^2\epsilon_{s-j}^2+o_{a.s.}(1)\,. \tag{4.20}$$

(c) *For* $0\le j,l<\infty$ *and* $j\ne l$, *then*

$$\frac{1}{t}\sum_{s=\max(j+2,l+2)}^{t}\kappa_j^{\phi}(s)\epsilon_{s-j}\kappa_l^{\phi}(s)\epsilon_{s-l}\xrightarrow{a.s.}0\,. \tag{4.21}$$

Proof of (a). By the boundedness of $\bar{P}_t^{-1},\phi_t,e_t$ (Proposition 4.7) and since $|\tilde{\kappa}_m(t)|\le\tilde{\kappa}_m$ for all $m\ge0$ and $t\ge1$,

$$\begin{aligned}
&\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{l=1}^{\min(j,s-1)}\tilde{\kappa}_{j-l}(s)\prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}\right)^{p}\\
&\quad\le\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{m=0}^{j}\tilde{\kappa}_j\sum_{l=1}^{\min(j,s-1)}\prod_{i=1}^{l}(s-i)^{-1}|\bar{P}_{s-i}^{-1}||\phi_{s-i-1}||e_{s-i}|\right)^{p}\\
&\quad\le K(\xi)\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{l=1}^{\min(j,s-1)}\prod_{i=1}^{l}(s-i)^{-1}\right)^{p}.
\end{aligned}$$

And since for all $j\ge0$, $p\ge1$,

$$\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{l=1}^{\min(j,s-1)}\prod_{i=1}^{l}(s-i)^{-1}\right)^{p}\le\frac{K}{t}\sum_{s=2}^{t}(s-\min(j,s-1))^{-p}\longrightarrow0\,,$$

(4.17) follows, as does (4.19), by the boundedness of ϵ_t. Similarly,

$$\begin{aligned}
&\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{l=1}^{\min(j,s-1)}\tilde{\kappa}_{j-l}(s)\sum_{i=0}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}\right)^{p}\\
&\quad\le K(\xi)\frac{1}{t}\sum_{s=2}^{t}\left(\sum_{l=1}^{\min(j,s-1)}\sum_{i=0}^{l}(s-i)^{-1}\right)^{p}\\
&\quad\le K(\xi)\frac{K}{t}\sum_{s=2}^{t}\left(\sum_{i=0}^{\min(j,s-1)}(s-i)^{-1}\right)^{p}\longrightarrow0\,,
\end{aligned} \tag{4.22}$$

and (4.18) follows. □

Proof of (b). From (4.12) and the recursion (3.1a) for θ_t, we have, for $s \geq j+2$,

$$\begin{aligned}
\kappa_j^x(s) &= \sum_{l=0}^{j}(-\beta)^l \kappa_{j-l} \prod_{i=1}^{l} \theta_{s-i} \\
&= \sum_{l=0}^{j}(-\beta)^l \kappa_{j-l} \prod_{i=1}^{l} \left(\theta_{s-i-1} + (s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}\right) \\
&= \sum_{l=0}^{j}(-\beta)^l \kappa_{j-l} \prod_{i=1}^{l} \theta_{s-i-1} + \sum_{l=0}^{j}(-\beta)^l \kappa_{j-l} \prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i} \\
&= \kappa_j^x(s-1) + w_j^x(s)\,.
\end{aligned} \tag{4.23}$$

where

$$w_j^x(s) = \sum_{l=0}^{j}(-\beta)^l \kappa_{j-l} \prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}, \tag{4.24}$$

Continuing, from (4.14) and (4.23)–(4.24), for $s \geq j+2$,

$$\begin{aligned}
\kappa_j^\phi(s) &= \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l} \theta_{s-i} \\
&= \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l} \left(\theta_{s-i-1} + (s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}\right) \\
&= \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l} \theta_{s-i-1} \\
&\quad + \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i} \\
&= \sum_{l=0}^{j}(-1)^l (\kappa_{j-l}^x(s-1) + w_{j-l}^x(s)) \prod_{i=1}^{l} \theta_{s-i-1} \\
&\quad + \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i} \\
&= \kappa_j^\phi(s-1) + w_j^\phi(s),
\end{aligned} \tag{4.25}$$

where from (4.24),

$$\begin{aligned}
w_j^\phi(s) &= \sum_{l=0}^{j}(-1)^l w_{j-l}^x(s) \prod_{i=1}^{l} \theta_{s-i-1} \\
&\quad + \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i} \\
&= \sum_{l=0}^{j}(-1)^l \sum_{m=0}^{j-l}(-\beta)^m \kappa_{j-l-m} \prod_{i=1}^{m}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i} \prod_{n=1}^{l} \theta_{s-n-1} \\
&\quad + \sum_{l=0}^{j}(-1)^l \kappa_{j-l}^x(s) \prod_{i=1}^{l}(s-i)^{-1}\bar{P}_{s-i}^{-1}\phi_{s-i-1}e_{s-i}.
\end{aligned} \tag{4.26}$$

By (4.19) and (4.25)–(4.26),

$$\frac{1}{t}\sum_{s=j+2}^{t}(\kappa_j^{\phi}(s))^2\epsilon_{s-j}^2 = \frac{1}{t}\sum_{s=j+2}^{t}\left((\kappa_j^{\phi}(s-1))^2 + 2\kappa_j^{\phi}(s-1)w_j^{\phi}(s) + (w_j^{\phi}(s))^2\right)\epsilon_{s-j}^2 .$$

Applying an argument similar to that used for part (a), it follows by the boundedness of β and θ_t and the Cauchy-Schwarz inequality that $t^{-1}\sum_{s=j+2}^{t}(2\kappa_j^{\phi}(s-1)w_j^{\phi}(s) + (w_j^{\phi}(s))^2)\epsilon_{s-j}^2 = o_{a.s.}(1)$. Hence,

$$\frac{1}{t}\sum_{s=j+2}^{t}(\kappa_j^{\phi}(s))^2\epsilon_{s-j}^2 = \frac{1}{t}\sum_{s=j+2}^{t}(\kappa_j^{\phi}(s-1))^2\epsilon_{s-j}^2 + o_{a.s.}(1) .$$

Finally, since j is finite, then for $i \leq j$, it follows by applying the recursion (4.25) in $\kappa_j^{\phi}(t)$ $i-1$ additional times that (4.20) holds, because a finite sum of $o_{a.s.}(1)$ terms is $o_{a.s.}(1)$. □

Proof of (c). By parts (a) and (b), for $j \neq l$,

$$\begin{aligned}
&\frac{1}{t}\sum_{s=\max(j+2,l+2)}^{t}\kappa_j^{\phi}(s)\epsilon_{s-j}\kappa_l^{\phi}(s)\epsilon_{s-l} \\
&\quad= \frac{1}{t}\sum_{s=\max(j+2,l+2)}^{t}\Bigg\{\Bigg(\kappa_j^{\phi}(s-1) + \sum_{p=0}^{\min(j,s-1)}(-1)^p\kappa_{j-p}^{x}(s) \\
&\qquad\qquad\times\prod_{q=1}^{l}(s-q)^{-1}\bar{P}_{s-q}^{-1}\phi_{s-q-1}e_{s-q}\Bigg) \\
&\qquad\times\Bigg(\kappa_l^{\phi}(s-1) + \sum_{r=0}^{\min(j,s-1)}(-1)^r\kappa_{j-r}^{x}(s) \\
&\qquad\qquad\times\prod_{m=1}^{r}(s-m)^{-1}\bar{P}_{s-m}^{-1}\phi_{s-m-1}e_{s-m}\Bigg)\epsilon_{s-j}\epsilon_{s-l}\Bigg\} \\
&\quad= \frac{1}{t}\sum_{s=\max(j+2,l+2)}^{t}\kappa_j^{\phi}(s-1)\kappa_l^{\phi}(s-1)\epsilon_{s-j}\epsilon_{s-l} + o_{a.s.}(1).
\end{aligned} \tag{4.27}$$

Without loss of generality, suppose $j < l < \infty$. From parts (a)–(b) and applying the argument that led to (4.27) $j-1$ additional times, we have that

$$\begin{aligned}
\frac{1}{t}\sum_{s=1}^{t}\kappa_j^{\phi}(s)\kappa_l^{\phi}(s)\epsilon_{s-j}\epsilon_{s-l} &= \frac{1}{t}\sum_{s=j+1}^{t}\kappa_j^{\phi}(s-j)\kappa_l^{\phi}(s-j)\epsilon_{s-j}\epsilon_{s-l} + o_{a.s.}(1) \\
&= \frac{1}{t}\sum_{s=1}^{t-j}\kappa_j^{\phi}(s)\kappa_l^{\phi}(s)\epsilon_{s-(l-j)}\epsilon_s + o_{a.s.}(1) , \\
&= \frac{1}{t}\sum_{s=1}^{t}\kappa_j^{\phi}(s)\kappa_l^{\phi}(s)\epsilon_{s-(l-j)}\epsilon_s + o_{a.s.}(1) ,
\end{aligned}$$

since by (D4) and the fact that $|\kappa_m^{\phi}(t)| \leq K(\xi) < \infty$ for all $m \geq 0$,

$$t^{-1}\sum_{s=t-j+1}^{t}\kappa_j^{\phi}(s)\kappa_l^{\phi}(s)\epsilon_{s-(l-j)}\epsilon_s = o_{a.s.}(1) .$$

Since $j < l$, $\epsilon_{s-(l-j)}$ is $\mathcal{F}_{s-1}$-measurable, as the σ-fields are increasing. Set $\tilde{z}_s = \kappa_j^\phi(s)\kappa_l^\phi(s)\epsilon_{s-(l-j)}$, which is $\mathcal{F}_{s-1}$-measurable by part (b) of Proposition 4.7. Then from boundedness, $\limsup \frac{1}{t}\sum_{s=1}^t \tilde{z}_s^2 \le (\sup_t |\kappa^\phi(t)|)^4(\sup_t |\epsilon_t|)^2 < \infty$, and thus from Proposition 4.2, $t^{-1}\sum_{s=1}^t \tilde{z}_s\epsilon_s \xrightarrow{a.s.} 0$ and therefore (4.21) holds and the proof of the Lemma is complete. □

Lemma 4.2. *For each $u \ge 0$, under the assumptions of Theorem 4.1,*

$$\frac{1}{t}\sum_{s=1}^t \phi_s^2 = \frac{1}{t}\sum_{s=1}^t \left(\sum_{j=0}^u \kappa_j^\phi(s)\epsilon_{s-j}\right)^2 + r_1(t,u) \tag{4.28}$$

where $\lim_u \limsup_t |r_1(t,u)| = 0$.

Proof.

$$\begin{aligned}\frac{1}{t}\sum_{s=1}^t \phi_s^2 &= \frac{1}{t}\sum_{s=1}^t \left(\sum_{j=0}^\infty \kappa_j^\phi(s)\epsilon_{s-j}\right)^2 \\ &= \frac{1}{t}\sum_{s=1}^t \left(\sum_{j=0}^u \kappa_j^\phi(s)\epsilon_{s-j}\right)^2 + \frac{2}{t}\sum_{s=1}^t \left(\sum_{j=0}^u \kappa_j^\phi(s)\epsilon_{s-j}\sum_{l=u+1}^\infty \kappa_l^\phi(s)\epsilon_{s-l}\right) \\ &\quad + \frac{1}{t}\sum_{s=1}^t \left(\sum_{j=u+1}^\infty \kappa_j^\phi(s)\epsilon_{s-j}\right)^2 .\end{aligned}$$

Let $r_1(t,u)$ be the sum of the last two terms. Recall from Proposition 4.7 and (4.15) that $|\kappa_j^\phi(t)| \le \kappa_j^\phi$ where $\sum_{j=0}^\infty \kappa_j^\phi < \infty$ a.s. From this and (D4), it follows that

$$\lim_u \limsup_t |r_1(t,u)| \le K(\xi)\lim_u \left\{\sum_{j=0}^u \kappa_j^\phi \sum_{l=u+1}^\infty \kappa_l^\phi + \left(\sum_{j=u+1}^\infty \kappa_j^\phi\right)^2\right\} = 0\,,$$

and consequently that (4.28) holds. □

Lemma 4.3. *For each $u \ge 0$, under the assumptions of Theorem 4.1,*

$$\frac{1}{t}\sum_{s=1}^t \left(\sum_{j=0}^u \kappa_j^\phi(s)\epsilon_{s-j}\right)^2 = \frac{\sigma_\epsilon^2}{t}\sum_{s=1}^t\sum_{j=0}^u \left(\kappa_j^\phi(s)\right)^2 + o_{a.s.}(1). \tag{4.29}$$

Proof.

$$\frac{1}{t}\sum_{s=1}^t \left(\sum_{j=0}^u \kappa_j^\phi(s)\epsilon_{s-j}\right)^2 = \frac{1}{t}\sum_{s=1}^t\sum_{j=0}^u (\kappa_j^\phi(s))^2\epsilon_{s-j}^2 + \frac{1}{t}\sum_{s=1}^t\sum_{j\ne l}^u \kappa_j^\phi(s)\epsilon_{s-j}\kappa_l^\phi(s)\epsilon_{s-l}\,.$$

Since u is finite, by Lemma 4.1, part (c),

$$\frac{1}{t}\sum_{s=1}^t\sum_{j\ne l}^u \kappa_j^\phi(s)\epsilon_{s-j}\kappa_l^\phi(s)\epsilon_{s-l} = o_{a.s.}(1)\,,$$

and so it remains to consider $t^{-1}\sum_{s=1}^{t}\sum_{j=0}^{u}(\kappa_j^{\phi}(s))^2\epsilon_{s-j}^2$. Consider the martingale difference sequence $\tilde{\epsilon}_t = \epsilon_t^2 - E[\epsilon_t^2|\mathcal{F}_{t-1}] = \epsilon_t^2 - \sigma_\epsilon^2$ (recall that $E[\epsilon_t^2|\mathcal{F}_{t-1}] = \sigma_\epsilon^2$). From (D4), $\tilde{\epsilon}_t$ is bounded a.s., hence $\sup_{-\infty<t<\infty} E[|\tilde{\epsilon}_t|^p|\mathcal{F}_{t-1}^{\epsilon}] < \infty$ a.s., so we can apply Proposition 4.2 to $\tilde{\epsilon}_t$.

For any $j \le s$, consider $t^{-1}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2\tilde{\epsilon}_s$. Since $\limsup t^{-1}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2 < \infty$, by Proposition 4.2, $t^{-1}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2\tilde{\epsilon}_s \xrightarrow{a.s.} 0$, hence, $t^{-1}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2(\epsilon_s^2 - \sigma_\epsilon^2) \xrightarrow{a.s.} 0$. By an argument like that used to prove part (b) of Lemma 4.1, it follows that

$$\begin{aligned}
\frac{1}{t}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2\tilde{\epsilon}_{s-j} &= \frac{1}{t}\sum_{s=j+1}^{t}(\kappa_j^{\phi}(s-j))^2\tilde{\epsilon}_{s-j} + o_{a.s.}(1)\,,\\
&= \frac{1}{t}\sum_{s=1}^{t-j}(\kappa_j^{\phi}(s))^2\tilde{\epsilon}_{s} + o_{a.s.}(1)\,,\\
&= \frac{1}{t}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2\tilde{\epsilon}_{s} + o_{a.s.}(1) \ \text{(since } j \le s),\\
&= \frac{1}{t}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2\tilde{\epsilon}_{s} + o_{a.s.}(1) = o_{a.s.}(1)\,,
\end{aligned}$$

i.e., $t^{-1}\sum_{s=1}^{t}(\kappa_j^{\phi}(s))^2(\epsilon_{s-j}^2 - \sigma_\epsilon^2) \xrightarrow{a.s.} 0$ for all $j \le u$. Finally, since u is finite, $t^{-1}\sum_{s=1}^{t}\sum_{j=0}^{u}(\kappa_j^{\phi}(s))^2(\epsilon_{s-j}^2 - \sigma_\epsilon^2) \xrightarrow{a.s.} 0$, and (4.29) holds and the proof of the Lemma is complete. □

Lemma 4.4. *Under the assumptions of Theorem 4.1, for each $u \ge 0$ and $0 \le k^* < \infty$, we have*

$$\begin{aligned}
\frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}\sum_{j=0}^{u}\left(\kappa_j^{\phi}(s)\right)^2 = \frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}\sum_{j=0}^{u}\left(\sum_{l=0}^{j}(-\theta_{s+k^*})^l\sum_{p=0}^{j-l}(-\beta\theta_{s+k^*})^p\kappa_{j-l-p}\right)^2\\
+ o_{a.s.}(1).
\end{aligned} \tag{4.30}$$

Proof. First suppose $k^* = 0$. Recalling from (4.12) and (4.14), for $s \ge j+1$,

$$\begin{aligned}
\kappa_j^{\phi}(s) &= \sum_{l=0}^{j}(-1)^l\kappa_{j-l}^{x}(s)\prod_{i=1}^{l}\theta_{s-i}\\
&= \sum_{l=0}^{j}(-1)^l\left(\sum_{p=0}^{j-l}(-\beta)^p\kappa_{j-l-p}\prod_{r=1}^{p}\theta_{s-r}\right)\prod_{i=1}^{l}\theta_{s-i}.
\end{aligned} \tag{4.31}$$

From (3.2) and (4.31), it follows that for $s \ge j+1$,

$$\begin{aligned}
\kappa_j^{\phi}(s) &= \sum_{l=0}^{j}(-1)^l\kappa_{j-l}^{x}(s)\prod_{i=1}^{l}\theta_{s-i} = \sum_{l=0}^{j}(-1)^l\theta_{s-1}\kappa_{j-l}^{x}(s)\prod_{i=2}^{l}\theta_{s-i}\\
&= \sum_{l=0}^{j}(-1)^l(\theta_s - s^{-1}\bar{P}_s^{-1}\phi_{s-1}e_s)\kappa_{j-l}^{x}(s)\prod_{i=2}^{l}\theta_{s-i}
\end{aligned} \tag{4.32}$$

$$= \sum_{l=0}^{j}(-1)^l \theta_s \kappa^x_{j-l}(s) \prod_{i=2}^{l} \theta_{s-i} - \sum_{l=0}^{j}(-1)^l s^{-1} \bar{P}_s^{-1} \phi_{s-1} e_s \kappa^x_{j-l}(s) \prod_{i=2}^{l} \theta_{s-i}.$$

Next, taking the square of (4.32), we obtain

$$\begin{aligned}(\kappa^\phi_j(s))^2 = &\left(\sum_{l=0}^{j}(-1)^l \theta_s \kappa^x_{j-l}(s) \prod_{i=2}^{l} \theta_{s-i}\right)^2 \\ &- 2\sum_{l=0}^{j}(-1)^l \\ &\times \sum_{m=0}^{j}\left\{(-1)^m (\theta_s \kappa^x_{j-l}(s) \left(\prod_{i=2}^{l} \theta_{s-i}\right) s^{-1} \bar{P}_s^{-1} \phi_{s-1} e_s) \kappa^x_{j-m}(s) \prod_{p=2}^{m} \theta_{s-p}\right\} \\ &+ \left(\sum_{l=0}^{j}(-1)^l s^{-1} \bar{P}_s^{-1} \phi_{s-1} e_s \kappa^x_{j-l}(s) \prod_{i=2}^{l} \theta_{s-i}\right)^2,\end{aligned} \tag{4.33}$$

and from the boundedness of θ_t and an argument like that used to prove (4.17), it follows that

$$\frac{\sigma_\epsilon^2}{t} \sum_{s=j+1}^{t} (\kappa^\phi_j(s))^2 = \frac{\sigma_\epsilon^2}{t} \sum_{s=j+1}^{t} \left(\sum_{l=0}^{j}(-1)^l \theta_s \kappa^x_{j-l}(s) \prod_{i=2}^{l} \theta_{s-i}\right)^2 + o_{a.s.}(1)\,. \tag{4.34}$$

Consider next the r.h.s. of (4.34). From (3.2),

$$\begin{aligned}&\sum_{l=0}^{j}(-1)^l \theta_s \kappa^x_{j-l}(s) \prod_{i=2}^{l} \theta_{s-i} \\ &\quad = \sum_{l=0}^{j}(-1)^l \theta_s \theta_{s-2} \kappa^x_{j-l}(s) \prod_{i=3}^{l} \theta_{s-i} \\ &\quad = \sum_{l=0}^{j}(-1)^l \theta_s \left(\theta_s - \sum_{m=0}^{1}(s-m)^{-1} \bar{P}_{s-m}^{-1} \phi_{s-m-1} e_{s-m}\right) \kappa^x_{j-l}(s) \prod_{i=3}^{l} \theta_{s-i} \\ &\quad = \sum_{l=0}^{j}(-1)^l \theta_s^2 \kappa^x_{j-l}(s) \prod_{i=3}^{l} \theta_{s-i} \\ &\qquad - \sum_{l=0}^{j}(-1)^l \theta_s \sum_{m=0}^{1}(s-m)^{-1} \bar{P}_{s-m}^{-1} \phi_{s-m-1} e_{s-m} \kappa^x_{j-l}(s) \prod_{i=3}^{l} \theta_{s-i}.\end{aligned}$$

Therefore, again from boundedness of θ_t and an argument like that used to prove (4.17),

$$\frac{\sigma_\epsilon^2}{t} \sum_{s=j+1}^{t} (\kappa^\phi_j(s))^2 = \frac{\sigma_\epsilon^2}{t} \sum_{s=j+1}^{t} \left(\sum_{l=0}^{j}(-1)^l \theta_s^2 \kappa^x_{j-l}(s) \prod_{i=3}^{l} \theta_{s-i}\right)^2 + o_{a.s.}(1)\,.$$

Applying the argument $l-2$ additional times, it follows that

$$\frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}(\kappa_j^\phi(s))^2 = \frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}\left(\sum_{l=0}^{j}(-\theta_s)^l\kappa_{j-l}^x(s)\right)^2 + o_{a.s.}(1)\,. \tag{4.35}$$

Next working on the r.h.s. of (4.35), from (4.12):

$$\begin{aligned}\sum_{l=0}^{j}(-\theta_s)^l\kappa_{j-l}^x(s) &= \sum_{l=0}^{j}(-\theta_s)^l\left(\sum_{p=0}^{j-l}(-\beta)^p\kappa_{j-l-p}\prod_{r=1}^{p}\theta_{s-r}\right)\\ &= \sum_{l=0}^{j}(-\theta_s)^l\left(\sum_{p=0}^{j-l}(-\beta)^p(\theta_s - s^{-1}\bar{P}_s^{-1}\phi_{s-1}e_s)\kappa_{j-l-p}\prod_{r=2}^{p}\theta_{s-r}\right)\\ &= \sum_{l=0}^{j}(-\theta_s)^l\left(\sum_{p=0}^{j-l}(-\beta)^p\theta_s\kappa_{j-l-p}\prod_{r=2}^{p}\theta_{s-r}\right)\\ &\quad - \sum_{l=0}^{j}(-\theta_s)^l\left(\sum_{p=0}^{j-l}(-\beta)^p s^{-1}\bar{P}_s^{-1}\phi_{s-1}e_s\kappa_{j-l-p}\prod_{r=2}^{p}\theta_{s-r}\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}\left(\sum_{l=0}^{j}(-\theta_s)^l\kappa_{j-l}^x(s)\right)^2\\ &\quad= \frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}\left(\sum_{l=0}^{j}(-\theta_s)^l\sum_{p=0}^{j-l}(-\beta)^p\theta_s\kappa_{j-l-p}\prod_{r=2}^{p}\theta_{s-r}\right)^2 + o_{a.s.}(1).\end{aligned}$$

Applying the argument $p-1$ additional times, it follows that

$$\begin{aligned}\frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}\left(\sum_{l=0}^{j}(-\theta_s)^l\kappa_{j-l}^x(s)\right)^2 &= \frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}\left(\sum_{l=0}^{j}(-\theta_s)^l\sum_{p=0}^{j-l}(-\beta\theta_s)^p\kappa_{j-l-p}\right)^2\\ &\quad + o_{a.s.}(1)\,,\end{aligned}$$

and, since j is finite,

$$\frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}(\kappa_j^\phi(s))^2 = \frac{\sigma_\epsilon^2}{t}\sum_{s=j+1}^{t}\left(\sum_{l=0}^{j}(-\theta_s)^l\sum_{p=0}^{j-l}(-\beta\theta_s)^p\kappa_{j-l-p}\right)^2 + o_{a.s.}(1)\,.$$

Finally, since u is finite, (4.30) follows for $k^*=0$.

From (3.2), for any finite $k^*>0$,

$$\theta_{s+k^*} = \theta_s + \sum_{r=0}^{k^*-1}(s+k^*-r)^{-1}\bar{P}_{s+k^*-r}^{-1}\phi_{s+k^*-r-1}e_{s+k^*-r}\,.$$

Set

$$\lambda(s,k^*) = \sum_{r=0}^{k^*-1}(s+k^*-r)^{-1}\bar{P}_{s+k^*-r}^{-1}\phi_{s+k^*-r-1}e_{s+k^*-r}.$$

For every integer $l \geq 0$, the binomial formula yields

$$\theta_{s+k^*}^l = \theta_s^l + \binom{l}{1} \theta_{s+k^*}^{l-1} \lambda(s,k^*) + \cdots + \binom{l}{l-1} \theta_{s+k^*} \lambda^{l-1}(s,k^*) + \lambda^l(s,k^*)\,.$$

Substituting this result into the r.h.s. of

$$\frac{\sigma_\epsilon^2}{t} \sum_{s=1}^{t} (\kappa_j^\phi(s))^2 = \frac{\sigma_\epsilon^2}{t} \sum_{s=j+1}^{t} \left(\sum_{l=0}^{j} (-\theta_{s+k^*})^l \sum_{p=0}^{j-l} (-\beta\theta_{s+k^*})^p \kappa_{j-l-p} \right)^2 + o_{a.s.}(1)\,,$$

which follows from Lemma 4.1, and noting that each resulting term involving $\lambda(s,k^*)$ is $o_{a.s.}(1)$ by (4.18), the proof of (4.30) and of the Lemma is reduced to the result just established for $k^* = 0$. □

Lemma 4.5. *Under the assumptions of Theorem 4.1, for any finite u,*

$$\frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \kappa_j^e(s) \kappa_{j-1}^\phi(s-1) \epsilon_{s-j}^2 = \frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \kappa_j^e(s) \kappa_{j-1}^\phi(s-1) + o_{a.s.}(1). \tag{4.36}$$

Proof. Since u is finite and for any finite $j \leq u$, $\limsup t^{-1}\sigma_\epsilon^2 | \sum_{s=2}^{t} \kappa_j^e(s) \times \kappa_{j-1}^\phi(s-1)| < \infty$, then the result follows by an argument similar to that used to prove part (c) of Lemma 4.1. □

Proposition 4.8. *Under the assumptions of Theorem 4.1, the sequence $\{\hat{\theta}_t\}$ defined by (4.1) satisfies $\theta_t - \hat{\theta}_t = o_{a.s.}(1)$.*

Proof. For simplicity, first assume that $k^* = 0$; i.e., $|\theta_t| \leq K^* < 1$ for all t. From the results of Proposition 4.7 and Lemmas 4.2–4.4, for any $u < \infty$:

$$\begin{aligned}
\frac{1}{t} \sum_{s=1}^{t} \phi_s^2 &= \frac{1}{t} \sum_{s=1}^{t} \left(\sum_{j=0}^{u} \kappa_j^\phi(s) \epsilon_{s-j} \right)^2 + r_1(t,u) \ \text{(Lemma 4.2)} \\
&= \frac{\sigma_\epsilon^2}{t} \sum_{s=1}^{t} \sum_{j=0}^{u} \left(\kappa_j^\phi(s) \right)^2 + o_{a.s.}(1) + r_1(t,u) \ \text{(Lemma 4.3)} \\
&= \frac{\sigma_\epsilon^2}{t} \sum_{s=1}^{t} \sum_{j=0}^{u} \left(\sum_{l=0}^{j} (-1)^l \prod_{i=1}^{l} \theta_{s-i} \sum_{p=0}^{j-l} (-\beta)^p \kappa_{j-l-p} \prod_{r=1}^{p} \theta_{s-r} \right)^2 + \ o_{a.s.}(1) \\
&\quad + r_1(t,u) \\
&= \frac{\sigma_\epsilon^2}{t} \sum_{s=1}^{t} \sum_{j=0}^{u} \left(\sum_{l=0}^{j} (-\theta_s)^l \sum_{p=0}^{j-l} (-\beta\theta_s)^p \kappa_{j-l-p} \right)^2 + \ o_{a.s.}(1) \\
&\quad + r_1(t,u) \ \text{(Lemma 4.4)}
\end{aligned}$$

where $\lim_u \limsup_t |r_1(t,u)| = 0$. By (2.3), Parseval's relation and convolution [22, pp. 61-66], it follows that

$$\begin{aligned}
&\frac{\sigma_\epsilon^2}{t} \sum_{s=1}^{t} \sum_{j=0}^{\infty} \left(\sum_{l=0}^{j} (-\theta_s)^l \sum_{p=0}^{j-l} (-\beta\theta_s)^p \kappa_{j-l-p} \right)^2 \\
&\qquad = \frac{1}{t} \sum_{s=1}^{t} \int_{-\pi}^{\pi} \frac{1}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2} g_y(\omega) d\omega\,,
\end{aligned}$$

and so

$$\frac{1}{t}\sum_{s=1}^{t}\phi_s^2 = \frac{1}{t}\sum_{s=1}^{t}\int_{-\pi}^{\pi}\frac{1}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2}g_y(\omega)d\omega + r_2(t,u)$$

where

$$r_2(t,u) = r_1(t,u) + \frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}\sum_{j=u+1}^{\infty}\left(\sum_{l=0}^{j}(-\theta_s)^l\sum_{p=0}^{j-l}(-\beta\theta_s)^p\kappa_{j-l-p}\right)^2.$$

Since $|\theta_t| \le K^* < 1$, it follows that

$$\begin{aligned}&\frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}\sum_{j=u+1}^{\infty}\left(\sum_{l=0}^{j}(-\theta_s)^l\sum_{p=0}^{j-l}(-\beta\theta_s)^p\kappa_{j-l-p}\right)^2\\ &\qquad\le \frac{\sigma_\epsilon^2}{t}\sum_{j=u+1}^{\infty}\left(\sum_{l=0}^{j}(K^*)^l\sum_{p=0}^{j-l}(\beta K^*)^p\kappa_{j-l-p}\right)^2 \overset{u\to\infty}{\longrightarrow} 0\end{aligned}$$

because $\sum_{j=0}^{\infty}\left(\sum_{l=0}^{j}(K^*)^l\sum_{p=0}^{j-l}(\beta K^*)^p\kappa_{j-l-p}\right)^2 < \infty$. Hence,

$$\lim_u \limsup_t \frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}\sum_{j=u+1}^{\infty}\left(\sum_{l=0}^{j}(-\theta_s)^l\sum_{p=0}^{j-l}(-\beta\theta_s)^p\kappa_{j-l-p}\right)^2 = 0\,,$$

and consequently, $\lim_u \limsup_t |r_2(t,u)| = 0$. It follows that

$$\left|\frac{1}{t}\sum_{s=1}^{t}\phi_s^2 - \frac{1}{t}\sum_{s=1}^{t}\int_{-\pi}^{\pi}\frac{1}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2}g_y(\omega)d\omega\right| \overset{a.s.}{\longrightarrow} 0\,. \tag{4.37}$$

Next, for $s \ge 2$ we use $\kappa_j^z(s) = \kappa_j^e(s) + \theta_{s-1}\kappa_{j-1}^\phi(s-1)$ from (4.16) to obtain that for any $u < \infty$,

$$\begin{aligned}\frac{1}{t}\sum_{s=2}^{t}z_s\phi_{s-1} &= \frac{1}{t}\sum_{s=2}^{t}\sum_{j=0}^{\infty}\kappa_j^z(s)\epsilon_{s-j}\sum_{l=0}^{\infty}\kappa_l^\phi(s-1)\epsilon_{s-1-l}\\ &= \frac{1}{t}\sum_{s=2}^{t}\sum_{j=0}^{u}\kappa_j^z(s)\epsilon_{s-j}\sum_{l=0}^{u}\kappa_l^\phi(s-1)\epsilon_{s-1-l} + r_3(t,u)\\ &= \frac{1}{t}\sum_{s=2}^{t}\sum_{j=1}^{u}\kappa_j^z(s)\kappa_{j-1}^\phi(s-1)\epsilon_{s-j}^2 + o_{a.s.}(1) + r_3(t,u)\\ &= \frac{1}{t}\sum_{s=2}^{t}\sum_{j=1}^{u}\left\{\left(\kappa_j^e(s) + \theta_{s-1}\kappa_{j-1}^\phi(s-1)\right)\kappa_{j-1}^\phi(s-1)\right\}\epsilon_{s-j}^2\\ &\quad + o_{a.s.}(1) + r_3(t,u)\\ &= \frac{1}{t}\sum_{s=2}^{t}\sum_{j=1}^{u}\kappa_j^e(s)\kappa_{j-1}^\phi(s-1)\epsilon_{s-j}^2\\ &\quad + \frac{1}{t}\sum_{s=2}^{t}\sum_{j=1}^{u}\theta_{s-1}\left(\kappa_{j-1}^\phi(s-1)\right)^2\epsilon_{s-j}^2 + o_{a.s.}(1) + r_3(t,u)\end{aligned} \tag{4.38}$$

$$= \frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \kappa_j^e(s) \kappa_{j-1}^\phi(s-1) \text{ (Lemma 4.5)}$$
$$+ \frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \theta_{s-1} \left(\kappa_{j-1}^\phi(s-1) \right)^2 + o_{a.s.}(1) + r_3(t,u) \,,$$

where

$$r_3(t,u) = \frac{1}{t} \sum_{s=2}^{t} \left(\sum_{j=u+1}^{\infty} \kappa_j^z(s) \epsilon_{s-j} \sum_{l=0}^{u} \kappa_l^\phi(s-1) \epsilon_{s-1-l} \right)$$
$$+ \frac{1}{t} \sum_{s=2}^{t} \left(\sum_{j=0}^{u} \kappa_j^z(s) \epsilon_{s-j} \sum_{l=u+1}^{\infty} \kappa_l^\phi(s-1) \epsilon_{s-1-l} \right)$$
$$+ \frac{1}{t} \sum_{s=2}^{t} \left(\sum_{j=u+1}^{\infty} \kappa_j^z(s) \epsilon_{s-j} \sum_{l=u+1}^{\infty} \kappa_l^\phi(s-1) \epsilon_{s-1-l} \right) .$$

By an argument similar to that applied to $r_1(t,u)$ in the proof of Lemma 4.2, one obtains $\lim_u \limsup_t |r_3(t,u)| = 0$. As shown above, the second term of (4.38), $t^{-1}\sigma_\epsilon^2 \sum_{s=2}^{t} \sum_{j=1}^{u} \theta_{s-1} \left(\kappa_{j-1}^\phi(s-1) \right)^2$, is equal to

$$\frac{1}{t} \sum_{s=1}^{t} \int_{-\pi}^{\pi} \frac{\theta_s}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2} g_y(\omega) d\omega + r_4(t,u)$$

with

$$r_4(t,u) = t^{-1}\sigma_\epsilon^2 \sum_{s=2}^{t} \sum_{j=u+1}^{\infty} \theta_{s-1} \left(\kappa_{j-1}^\phi(s-1) \right)^2 + o_{a.s.}(1)$$

and $\lim_u \limsup_t |r_4(t,u)| = 0$. Hence, it remains to consider the first term of (4.38). From (4.11) and (4.31),

$$\frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \kappa_j^e(s) \kappa_{j-1}^\phi(s-1)$$
$$= \frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \left\{ \sum_{l=0}^{j} (-1)^l \kappa_{j-l} \prod_{i=1}^{l} \theta_{s-i} \right\}$$
$$\times \left\{ \sum_{m=0}^{j-1} (-1)^m \prod_{p=1}^{m} \theta_{s-1-p} \left(\sum_{n=0}^{j-l-1} (-\beta)^n \kappa_{j-l-1-n} \prod_{r=1}^{n} \theta_{s-1-r} \right) \right\}$$
$$= \frac{\sigma_\epsilon^2}{t} \sum_{s=2}^{t} \sum_{j=1}^{u} \left\{ \left(\sum_{l=0}^{j} (-\theta_s)^l \kappa_{j-l} \right) \left(\sum_{m=0}^{j-1} (-\theta_s)^m \sum_{n=0}^{j-l-1} (-\beta\theta_s)^n \kappa_{j-l-1-n} \right) \right\}$$
$$+ o_{a.s.}(1) \,,$$

and, since again by (2.3), Parseval's relation and convolution,

$$\sigma_\epsilon^2 \sum_{j=1}^{\infty} \left\{ \left(\sum_{l=0}^{j} (-\theta_s)^l \kappa_{j-l} \right) \left(\sum_{m=0}^{j-1} (-\theta_s)^m \sum_{n=0}^{j-l-1} (-\beta\theta_s)^n \kappa_{j-l-1-n} \right) \right\}$$
$$= \int_{-\pi}^{\pi} \frac{1}{(1+\theta_s e^{-i\omega})} \frac{e^{i\omega}}{(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})} g_y(\omega) d\omega \,,$$

the first term of (4.38), $t^{-1}\sigma_\epsilon^2 \sum_{s=2}^{t}\sum_{j=1}^{u} \kappa_j^e(s)\kappa_{j-1}^{\phi}(s-1)$, is equal to

$$\frac{1}{t}\sum_{s=1}^{t} \frac{1}{(1+\theta_s e^{-i\omega})} \frac{e^{i\omega}}{(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})} g_y(\omega)d\omega + r_5(t,u)$$

with

$$r_5(t,u) = \frac{\sigma_\epsilon^2}{t}\sum_{s=1}^{t}\sum_{j=u+1}^{\infty}\left\{\left(\sum_{l=0}^{j}(-\theta_s)^l\kappa_{j-l}\right) \times \left(\sum_{m=0}^{j-1}(-\theta_s)^m \sum_{n=0}^{j-l-1}(-\beta\theta_s)^n \kappa_{j-l-1-n}\right)\right\} + o_{a.s.}(1)\,.$$

An argument like that applied to $r_2(t,u)$ yields $\lim_u \limsup_t |r_5(t,u)| = 0$. Further, since

$$\begin{aligned}
&\int_{-\pi}^{\pi} \frac{1}{(1+\theta_s e^{-i\omega})} \frac{e^{i\omega}}{(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})} g_y(\omega)d\omega \\
&\quad + \int_{-\pi}^{\pi} \frac{\theta_s}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2} g_y(\omega)d\omega \\
&\quad = \int_{-\pi}^{\pi} \frac{e^{i\omega}(1+\beta\theta_s e^{-i\omega}) + \theta_s}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2} g_y(\omega)d\omega \\
&\quad = \int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_s}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2} g_y(\omega)d\omega\,,
\end{aligned}$$

we obtain

$$\left|\frac{1}{t}\sum_{s=2}^{t} z_s\phi_{s-1} - \frac{1}{t}\sum_{s=1}^{t}\int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_s}{|(1+\theta_s e^{i\omega})(1+\beta\theta_s e^{i\omega})|^2} g_y(\omega)d\omega\right| \xrightarrow{a.s.} 0\,. \tag{4.39}$$

Combining (4.37) and (4.39), it follows from (3.5) and (3.16) that $\theta_t - \hat{\theta}_t = o_{a.s.}(1)$ where $\hat{\theta}_t$ is given by (4.1). Finally, since k^* is finite, an argument similar to that used for Lemma 4.4 can be applied to show that (4.1) holds for the general case $k^* > 0$, completing the proofs of the proposition and part (a) of Theorem 4.1. □

Proposition 4.9. *Under the assumptions of Theorem 4.1 and with $t_0 = t_0(\xi)$ as in (a) of the Theorem, $\hat{\theta}_t$ defined by (4.1) satisfies the conditions of Proposition 4.6 for $\tilde{\Theta} = \Theta = (-1,1)$ for a Robbins-Monro recursion with $f(\theta) = f(\theta,\beta)$ as in (1.4).*

Proof. For $t \geq 2$, set

$$\tilde{P}_t = \sum_{s=1}^{t}\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{s+k^*}e^{i\omega})(1+\beta\theta_{s+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\,.$$

Then from (4.1),

$$\begin{aligned}
\hat{\theta}_t &= \tilde{P}_t^{-1}\sum_{s=1}^{t}\int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_{s+k^*}}{|(1+\theta_{s+k^*}e^{i\omega})(1+\beta\theta_{s+k^*}e^{i\omega})|^2} g_y(\omega)d\omega \\
&= \tilde{P}_t^{-1}\left\{\sum_{s=1}^{t-1}\int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_{s+k^*}}{|(1+\theta_{s+k^*}e^{i\omega})(1+\beta\theta_{s+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right. \\
&\quad \left. + \int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_{t+k^*}}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right\}
\end{aligned} \tag{4.40}$$

$$
\begin{aligned}
&= \tilde{P}_t^{-1}\left\{\tilde{P}_{t-1}\hat{\theta}_{t-1} + \int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_{t+k^*}}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right\} \\
&= \tilde{P}_t^{-1}\left\{\left(\tilde{P}_t - \int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right)\hat{\theta}_{t-1}\right. \\
&\quad \left. + \int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_{t+k^*}}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right\} \\
&= \hat{\theta}_{t-1} + \tilde{P}_t^{-1}\left\{\int_{-\pi}^{\pi} \frac{e^{i\omega} + (1+\beta)\theta_{t+k^*}}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right. \\
&\quad \left. - \int_{-\pi}^{\pi} \frac{\hat{\theta}_{t-1}}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right\} \\
&= \hat{\theta}_{t-1} + \tilde{P}_t^{-1}\int_{-\pi}^{\pi} \frac{e^{i\omega} + \beta\theta_{t+k^*}}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega \\
&\quad + \tilde{P}_t^{-1}(\theta_{t+k^*} - \hat{\theta}_{t-1})\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega \\
&= \hat{\theta}_{t-1} - \tilde{P}_t^{-1} f(\theta_{t+k^*}, \beta) \\
&\quad + \tilde{P}_t^{-1}(\theta_{t+k^*} - \hat{\theta}_{t-1})\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega \\
&= \hat{\theta}_{t-1} - \tilde{P}_t^{-1} f(\hat{\theta}_{t-1}, \beta) + \tilde{P}_t^{-1}\Big(f(\hat{\theta}_{t-1}, \beta) - f(\theta_{t+k^*}, \beta)\Big) \\
&\quad + \tilde{P}_t^{-1}(\theta_{t+k^*} - \hat{\theta}_{t-1})\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega \\
&= \hat{\theta}_{t-1} - \delta_t f(\hat{\theta}_{t-1}, \beta) + \delta_t \gamma_t\,,
\end{aligned}
$$

where, for $t \geq 2$,

$$
\delta_t = \tilde{P}_t^{-1} = \frac{1}{t}\left[\frac{1}{t}\sum_{s=1}^{t}\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{s+k^*}e^{i\omega})(1+\beta\theta_{s+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\right]^{-1}, \tag{4.41}
$$

and, for $t \geq t_0+1$ with t_0 as in (a) of Theorem 4.1 (which guarantees that $f(\hat{\theta}_{t-1}, \beta)$ below is finite),

$$
\begin{aligned}
\gamma_t &= \Big(f(\hat{\theta}_{t-1}, \beta) - f(\theta_{t+k^*}, \beta)\Big) \\
&\quad + (\theta_{t+k^*} - \hat{\theta}_{t-1})\int_{-\pi}^{\pi} \frac{1}{|(1+\theta_{t+k^*}e^{i\omega})(1+\beta\theta_{t+k^*}e^{i\omega})|^2} g_y(\omega)d\omega\,.
\end{aligned} \tag{4.42}
$$

For $|\theta| \leq K^* < 1$, it follows from (3.16) and (4.41), there exist finite, positive $\tilde{K}_1(\xi) \leq \tilde{K}_2(\xi)$ such that $0 < \tilde{K}_1(\xi) \leq t\,\delta_t \leq \tilde{K}_2(\xi) < \infty$. From this, it follows that $\delta_t \xrightarrow{a.s.} 0$ and $\sum_{s=1}^{t}\delta_s \geq \tilde{K}_1\sum_{s=1}^{t}k^{-1} \to \infty$. Next since k^* is finite, it follows from $\theta_{t-1} - \hat{\theta}_{t-1} = o_{a.s.}(1)$ (Proposition 4.7) that $\theta_{t+k^*} - \hat{\theta}_{t-1} = o_{a.s.}(1)$ and $f(\theta_{t+k^*}, \beta) - f(\hat{\theta}_{t-1}, \beta) = o_{a.s.}(1)$. Hence, $\gamma_t \xrightarrow{a.s.} 0$. The definition of t_0 in (a) of Theorem 4.1, guarantees that the remaining condition of Proposition 4.6 is satisfied, so the proposition is proved. □

We can now complete the proof of Theorem 4.1. By Proposition 4.6, $\hat{\theta}_t \xrightarrow{a.s.} \Theta_0^\beta$ and therefore also $\theta_t \xrightarrow{a.s.} \Theta_0^\beta$, which is compact by Proposition 4.4. Further, if y_t is an invertible ARMA process, then by Proposition 4.5, the set Θ_0^β is finite and

Proposition 4.6 shows θ_t converges on almost every realization to one of the finitely many $\theta \in \Theta_0^\beta$. Consequently, on the probability one event on which θ_t converges, its limit is a random variable θ with finitely many values. On the complementary event, θ can be defined to have any fixed value. This completes the proof of part (b) and with it the proof of the Theorem.

5. Discussion

The results obtained here provide a rigorous foundation for analyzing PLR and RML_2 for MA(1) models. An important conclusion from our results is that under misspecification, generally only RML_2 (i.e., the general algorithm with $\beta = 1$), not the simpler and more frequently considered PLR algorithm, can produce optimal coefficient estimates in the limit. In [5], Theorem 4.1 is applied to address convergence of PLR and convergence of RML_2 with a specific monitoring and modification scheme to ensure that iterates satisfy $|\theta_t| \leq K^* < 1$. In [5] we also provide a set of examples that show that the limits of θ_t from PLR and RML_2 can differ.

Ideas and techniques from the analysis of Hannan [12] of RML_2 for ARMA models played a key role in our analysis, particularly the idea of approximating the recursive algorithm's sequence by a sequence made more analyzable, replacing certain terms by their expected values, and replacing terms in an expression by finitely lagged values, as in our (4.20), so that martingale results like Propositions 4.1 and 4.2 can be applied. However we note that, because of a neglected $o_{a.s.}(1)$ term that depends on θ_t, Hannan did not actually establish that his auxiliary sequence, which we denote by $\tilde{\theta}_t$ to distinguish it from our $\hat{\theta}_t$, satisfies his (nonstandard) recursion scheme. Also, the convergence analysis he indicates for $\tilde{\theta}_t$, if its details could be verified, would only establish that the *limit inferior* of $\min_{\theta \in \Theta_0^1} |\tilde{\theta}_t - \theta|$ is zero a.s., see p. 773 of Hannan [12]. The stronger result with the *limit* is needed to establish convergence of the original recursive sequence to Θ_0^1. More information about problems we encountered with analyses in Hannan [12] can be found in [4, Appendix E].

The approximating sequence technique is similar to the Ordinary Differential Equation (ODE) method independently developed by Ljung [20] and Kushner [14, 15]. Specifically, the ODE method is a technique for providing asymptotic analysis of a time series (discrete stochastic process) via a deterministic continuous time stability analysis of a set of ODEs. For example, from the ODE method, Ljung makes convergence assertions for both PLR and RML_2 for ARMAX models including in the incorrect model situation [21]. Like Hannan, however, the analysis is incomplete. In the rigorous treatment of the ODE method presented by Benveniste et al [1] only the correct model situation is considered. Their results, however, do not apply to PLR or RML_2 [4, pp.65-67].

Clearly, the boundedness assumption (D4) is restrictive but it is typical in convergence analyses like ours. For example, boundedness is an explicit assumption in the deep correct model results obtained by Lai and Ying [17, equation (1.3)] as well as Ljung's ODE method assertions [21, condition S2, p.191] and is also required in the treatment by Benveniste et al. in which θ_t is assumed to be bounded to obtain verifiable conditions to prove asymptotic results [1, Theorem 15 and Corollary 16, p.238].

Finally, it is likely that Theorem 4.1 is generalizable to higher order moving average models and quite possibly ARMA models. However, to obtain convergence results, a multidimensional parameter vector θ version of Proposition 4.6 is needed.

The proof of Theorem 2.2.2 of [7] seems to provide the needed result if it can be shown that an appropriate Liapounov function exists for the vector-valued $f(\theta, \beta)$ associated with multidimensional θ for $0 \leq \beta < 1$. The generalization of the $\bar{L}(\theta)$ with vector θ provides the Liapounov function for the case $\beta = 1$.

Acknowledgment

The authors gratefully acknowledge the detailed review and insightful comments and suggestions of the referee.

References

[1] Benveniste, A., Métivier, M. and Priouret, P.(1990). *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, New York.

[2] Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226. MR323048

[3] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA. MR436499

[4] Cantor, J. L. (2001). Recursive and batch estimation for misspecified ARMA models. Ph.D. thesis, George Washington University.

[5] Cantor, J. L. and Findley, D. F. Recursive estimation of misspecified MA(1) models: Comparison of PLR and RML_2. In preparation.

[6] Chen, H. and Guo, L. (1991). *Identification and Stochastic Adaptive Control*. Birkhäuser, Boston, MA.

[7] Chen, H.-F. (2002). *Stochastic Approximation and Its Applications*. Nonconvex Optimization and its Applications, Vol. **64**. Kluwer Academic Publishers, Dordrecht. MR1942427

[8] Derevitskiĭ, D. P. and Fradkov, A. L. (1981). *Prikladnaya Teoriya Diskretnykh Adaptivnykh Sistem Upravleniya*. Nauka, Moscow. MR641849

[9] Findley, D. F. (2005). Convergence of a Robbins-Monro algorithm for recursive estimation with non-monotone weights and possibly multiple zeros. *Calcutta Statistical Association Bulletin* **56** (221) 1–16. Special 5th Triennial Proceedings Volume.

[10] Findley, D. F. and Wei, C. Z. (1993). Moment bounds for deriving time series CLTs and model selection procedures. *Statist. Sinica* **3** (2) 453–480. MR1243396

[11] Hannan, E. J. (1976). The convergence of some recursions. *Ann. Statist.* **4** (6) 1258–1270. MR519092

[12] Hannan, E. J. (1980). Recursive estimation based on ARMA models. *Ann. Statist.* **8** (4) 762–777. MR572620

[13] Henrici, P. (1974). *Applied and Computational Complex Analysis*. Wiley-Interscience [John Wiley & Sons], New York. MR372162

[14] Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Applied Mathematical Sciences, Vol. **26**. Springer-Verlag, New York. MR499560

[15] Kushner, H. J. and Yin, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. Applications of Mathematics (New York), Vol. **35**. Springer-Verlag, New York. MR1453116

[16] Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10** (1) 154–166. MR642726

[17] Lai, T. L. and Ying, Z. (1991). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM J. Control Optim.* **29** (5) 1061–1090. MR1110087

[18] Lai, T. L. and Ying, Z. (1992). Recursive solutions of estimating equations and adaptive spectral factorization. *IEEE Trans. Automat. Control* **37** (2) 240–243. MR1144904

[19] Lai, T. L. and Ying, Z. (2006). Efficient recursive estimation and adaptive control in stochastic regression and ARMAX models. *Statist. Sinica* **16** (3) 741–772.

[20] Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control* **AC-22** (4) 551–575. MR465458

[21] Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press Series in Signal Processing, Optimization, and Control, Vol. 4. MIT Press, Cambridge, MA. MR719192

[22] Oppenheim, A. and Schafer, R. (1975). *Digital Signal Processing*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

[23] Pötscher, B. M. (1987). Convergence results for maximum likelihood type estimators in multivariable ARMA models. *J. Multivariate Anal.* **21** (1) 29–52. MR877841

[24] Solo, V. (1979). The convergence of AML. *IEEE Transactions on Automatic Control* **AC-24** (6) 958–962.

[25] Zygmund, A. (1968). *Trigonometric Series*, Vols. I, II. Second edition, reprinted with corrections and some additions. Cambridge University Press, London. MR236587

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 48–59

DOI: 10.1214/074921706000000941

Estimation of AR and ARMA models by stochastic complexity

Ciprian Doru Giurcăneanu[1,*] and Jorma Rissanen[2,†,‡]

Tampere University of Technology, and Technical University of Tampere and Helsinki, and Helsinki Institute for Information Technology

Abstract: In this paper the stochastic complexity criterion is applied to estimation of the order in AR and ARMA models. The power of the criterion for short strings is illustrated by simulations. It requires an integral of the square root of Fisher information, which is done by Monte Carlo technique. The stochastic complexity, which is the negative logarithm of the Normalized Maximum Likelihood universal density function, is given. Also, exact asymptotic formulas for the Fisher information matrix are derived.

1. Introduction

The negative logarithm of the *NML* (Normalized Maximum Likelihood) universal model, called the *stochastic complexity*, provides a powerful criterion for estimation of the model structure such as the optimal collection of the regressor variables in the linear quadratic regression problem, [19], especially for small amounts of data. It involves the integral of the square root of the Fisher information, which is easy to calculate when the regressor matrix does not depend on the parameters. While modeling gaussian time series with AR models are instances of linear quadratic regression problems their order estimation poses trouble with the stochastic complexity for the reason that the regressor matrix is determined by the parameters, and the Fisher information is not constant. The same problem of course is also with the ARMA models, which have the additional difficulty of calculation of the maximum likelihood parameters.

In this paper we resort to Monte Carlo integration to overcome the problem posed by the nonconstant Fisher information and study by simulations the efficiency of the resulting order estimation criterion. Although exact formulas exist for the Fisher information matrix they are quite cumbersome to evaluate, and we consider asymptotic simplifications. This may run against the intent of getting a criterion for small amounts of data, but the asymptotic estimates appear to be good enough, and the resulting criterion for the short data sequences created is still superior among the competing criteria such as the *BIC* [20], which is equivalent with a crude asymptotic version of the *MDL* criterion [15], and a recently suggested one, *KICC* [21], or bias corrected Kullback-Leibler criterion.

[1]Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland, e-mail: ciprian.giurcaneanu@tut.fi

[2]140 Teresita Way, Los Gatos, CA 95032, USA, e-mail: jrrissanen@yahoo.com

[*]The work of C.D. Giurcăneanu was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program (2006–2011).

[†]J. Rissanen is affiliated with the Technical Universities of Tampere and Helsinki, and Helsinki Institute for Information Technology, Finland.

[‡]Corresponding author.

AMS 2000 subject classifications: primary 62B10; secondary 91B70.

Keywords and phrases: minimum description length principle, Fisher information, normalized maximum likelihood universal model, Monte Carlo technique.

We describe below the *NML* model for AR and ARMA class of models, and discuss its optimality properties. We also derive in the Appendix the asymptotic form of the Fisher information matrix for the general ARMA class of models.

2. Normalized maximum likelihood model

We consider the ARMA model:

$$y_t + \sum_{i=1}^{n} a_i y_{t-i} = e_t + \sum_{j=1}^{m} b_j e_{t-j}, \tag{1}$$

where e_t is zero-mean white Gaussian noise of variance σ^2. The integers m, n are nonnegative, and all coefficients a_i and b_j are real-valued. We can equivalently write $y_t = \dfrac{B(q)}{A(q)} e_t$, where $B(q) = 1 + b_1 q^{-1} + \cdots + b_m q^{-m}$, $A(q) = 1 + a_1 q^{-1} + \cdots + a_n q^{-n}$, and q^{-1} is the unit delay operator. We will use the notation ARMA(n,m) for the class of the normal density functions $\{f(y^N;\theta)\}$ defined by such processes, where $\theta = (a_1, \ldots, a_n, b_1, \ldots, b_m, \sigma^2)$, the parameters ranging over a subset of $\Re^k$, where $k = n + m + 1$. Let $\hat{\theta}(y^N)$ denote the maximum likelihood estimates of the parameters θ.

In order to define the range of the parameters properly we need to consider another equivalent parametrization in terms of the roots of the two polynomials

$$\prod_{i=1}^{n}(1 - g_i q^{-1}) y_t = \prod_{j=1}^{m}(1 - h_j q^{-1}) e_t, \tag{2}$$

together with the noise variance σ^2. We denote by g_i the zeros of $A(q)$ and by h_j the zeros of $B(q)$. There are no repeated poles or zeros nor pole-zero cancellations. We specify in the Appendix exactly the further restrictions on the type of the zeros but for now let the same symbol θ denote the new parameters ranging over $\Theta \subset \Re^k$.

Consider the *NML* density function, [3],[18],

$$\hat{f}(y^N; n, m) = \frac{f(y^N; \hat{\theta}(y^N))}{C_{k,n}},$$

where

$$\begin{aligned} C_{k,n} &= \int_{x^N : \hat{\theta}(x^N) \in \Omega} f(x^N; \hat{\theta}(x^N)) dx^N \\ &= \int_{\hat{\theta} \in \Omega} g(\hat{\theta}; \hat{\theta}) d\hat{\theta}, \end{aligned}$$

and $g(\hat{\theta};\theta)$ denotes the density function on the statistic $\hat{\theta}$ induced by $f(y^N;\theta)$. In the equation above, we use the identity $f(x^N; \hat{\theta}(x^N), \theta) = f(x^N | \hat{\theta}(x^N); \theta) g(\hat{\theta}(x^N); \theta)$, that is integrated first over x^N at the point $\hat{\theta}(x^N) = \hat{\theta} = \theta$ kept fixed, which gives unity, and then over $\hat{\theta}$.

Under the main assumption that the convergence in distribution by the Central Limit Theorem applies to the ML estimates, the stochastic complexity, $L(y^N; n, m) = \ln 1/\hat{f}(y^N; n, m)$, is given by

$$L(y^N; n, m) = -\ln f(y^N; \hat{\theta}(y^N)) + \frac{k}{2} \ln \frac{N}{2\pi} + \ln \int_{\Theta} |\mathbf{J}(\theta)|^{1/2} \mathrm{d}\theta + o(1), \tag{3}$$

where Θ denotes the parameter space, and $\mathbf{J}(\theta)$ is the Fisher information matrix [18]. The rate of convergence o(1) is determined by the convergence of the ML estimates to the normal density function.

To get a criterion for the structure in general we ought to add the code length needed to encode the structure, but here we take the simple case where the structure consists of a few first coefficients of the ARMA model, whose code length is much shorter than the stochastic complexity and ignored. (If k is not small, we can use the estimate $L(k) = \ln k + 2 \ln \ln k$.)

The *NML* model has the following two optimality properties, which justify its name:

(1) It is the unique solution $\hat{f} = \hat{g} = \hat{q}$ to the following maxmin problem

$$\max_g \min_q E_g \log \frac{f(y^N; \hat{\theta}(y^N))}{q(y^N)},$$

where g and q range over any sets that include $\hat{f}$. Notice that the logarithm of the ratio is the difference between the ideal code length $\log 1/q$ and the unattainable lower bound for any code length in the ARMA class.

(2) If the data generating distribution g is restricted to the ARMA class, the mean of the stochastic complexity with respect to the model θ cannot be beaten by any model what so ever, except for θ in a set whose volume goes to zero as N grows.

3. Linear regression with constant regressor matrix

Before discussing the AR models we illustrate the stochastic complexity criterion for linear quadratic regression with constant Fisher information by comparing it with the *BIC* and the *KICC* criteria in a simple polynomial fitting problem for small amounts of data.

For linear regression with a constant regressor matrix $\mathbf{X} = \{x_{it}\}$ the stochastic complexity criterion takes the form, [19],

$$\min_{\gamma \in \Gamma} \{(N-k) \ln \hat{\tau} + k \ln \hat{R} + (N-k-1) \ln \frac{1}{n-k} - (k-1) \ln k\}.$$

The index $\gamma = i_1, \ldots, i_k$, consists of the indices of the rows $\bar{\mathbf{x}}_\mathbf{i}$ of the $k \times n$ regressor matrix included in the linear combination

$$y_t = \sum_{i \in \gamma} \beta_i \bar{x}_{it} + e_t, \ t = 1, \ldots, N,$$

$\hat{\tau}$ is the minimized squared error per symbol, and $\hat{R} = \frac{1}{n} \hat{\boldsymbol{\beta}}^\top \mathbf{X}_\gamma \mathbf{X}_\gamma^\top \hat{\boldsymbol{\beta}}$, where $\mathbf{X}_\gamma$ is the $k \times n$ submatrix of $\mathbf{X}$ consisting of the retained rows.

Notice that there are no hyper parameters defining the range of the parameters β_i and τ. They have been renormalized away.

Example 1. We discuss an example of polynomial fitting considered in [21] to investigate the performances of a model selection criterion called *KICC*. It is obtained by an application of a bias correction to KIC (Kullback Information Criterion), [6], and it is recommended to be used in linear regression problems when the sample size is small. The underlying signal is generated by a third-order polynomial model $\tilde{y} = x^3 - 0.5x^2 - 5x - 1.5$, where the points $x_1, \ldots, x_N$ are chosen to be uniformly

TABLE 1

Order estimation of the polynomial model in Example 1. The true order is $k = 3$. For each criterion, the probability of correct estimation of the order is computed from 10^5 runs. Also shown is the probability of overestimation of the polynomial order ($4 \leq \hat{k} \leq 10$). The probability of underestimation ($0 \leq \hat{k} \leq 2$) is almost zero for all analyzed criteria. The best result for each sample size N is represented with bold font.

Order	Criterion	Sample size(N)								
		25	30	40	50	60	70	80	90	100
$\hat{\mathbf{k}} = \mathbf{k}$	*NML*	**0.94**	**0.95**	**0.96**	**0.97**	**0.97**	**0.97**	**0.98**	**0.98**	**0.98**
	BIC	0.79	0.84	0.89	0.91	0.93	0.94	0.95	0.95	0.95
	KICC	0.93	0.92	0.91	0.91	0.90	0.90	0.90	0.90	0.89
$\hat{k} > k$	*NML*	0.06	0.05	0.04	0.03	0.03	0.03	0.02	0.02	0.02
	BIC	0.21	0.16	0.11	0.09	0.07	0.06	0.05	0.05	0.05
	KICC	0.07	0.08	0.09	0.09	0.10	0.10	0.10	0.10	0.11

distributed in $[-3, 3]$. The measurements $y_1, \ldots, y_N$ are obtained by addition to $\tilde{y}_i$ zero-mean white Gaussian noise, whose variance is selected such that the signal-to-noise ratio is SNR=10 dB. For each number of data points N, between 25 and 100, 10^5 different realizations are produced, to which polynomials of degree $0, 1, \ldots, 10$ are fitted with the least squares method.

The estimates of the order of the polynomial obtained with the *NML*, *BIC* and *KICC* criteria are in Table 1. We have restricted our investigations only to these three criteria, because in [21] *KICC* was shown to outperform other six estimation criteria for $N = 25$ and $N = 30$. We see in the table that *NML* criterion performs better than *BIC* and *KICC* in all the cases studied. Observe that the number of correct estimations produced by *KICC* generally declines when more measurements are available, while the *BIC* and the *NML* results improve with increasing N. For example, *KICC* compares favorable with *BIC* for $N = 25$, but the situation is reversed for $N = 100$.

4. AR models

The likelihood density function for an AR model is given by

$$f(y^N; \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y_t + a_1 y_{t-1} + \cdots + a_n y_{t-n})^2},$$

where we put $y_t = 0$ for $t < 1$. The maximized likelihood is $\dfrac{1}{(2\pi e\hat{\sigma}^2)^{N/2}}$, where $\hat{\sigma}^2$ is the minimized sum per symbol $\hat{\sigma}^2 = \dfrac{1}{N}\displaystyle\sum_{t=1}^{N}(y_t + \hat{a}_1 y_{t-1} + \cdots + \hat{a}_n y_{t-n})^2$. The *NML* criterion (3) has now the expression

$$L(y^N; n) = \frac{N}{2}\ln(2\pi e\hat{\sigma}^2) + \frac{n+1}{2}\ln\frac{N}{2\pi} + \ln\int_{\Theta} |\mathbf{J}(\theta)|^{1/2} \mathrm{d}\theta + o(1). \tag{4}$$

The Fisher information matrix is given by $\begin{bmatrix} \mathbf{R}_{zz} & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}$, where

$$\mathbf{R}_{zz} = \begin{bmatrix} r_0 & r_1 & \cdots & r_{n-1} \\ r_1 & r_0 & \cdots & r_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n-1} & r_{n-2} & \cdots & r_0 \end{bmatrix},$$

and $r_i = E[z_t z_{t-i}]$ denote the covariances of the process $z_t = y_t/\sigma$ [9, 10]. Applying the formula in [12] for the parameters transformation and the well-known Vieta's formulae, it is easy to calculate the Fisher information matrix for the parameter set given by the model poles $g = (g_1, g_2, \ldots, g_n)$ and the noise variance σ^2.

Remark in (4) that the integral term makes the most important difference between the expression for the stochastic complexity and the BIC criterion. The integral has a lot of structural information which BIC lacks, and it generally increases with n, because the determinant increases.

We note that the contribution of the σ^2 to the integral is decoupled by the contribution of the other parameters. Consequently we ignore for all the AR models the contribution of σ^2 because we do not have any "natural" finite limits for the range of σ^2. The constrain to have a stable model restricts the domain of the magnitudes of the poles to be a hypercube.

Apart from the AR(1) case for which the integral in (3) can be found in a closed form, $\int_{-1}^{1} \frac{1}{\sqrt{1-g^2}} \mathrm{d}g = \pi$, the evaluation of the integral will be done by the Monte Carlo technique. To be more precise we use Sobol' sequences [14] to perform the Monte Carlo integration for AR(n) models with $1 \leq n \leq 6$. For these values all poles are complex if n is even, and exactly one pole is real-valued if n is odd, which can be taken advantage of in calculating the form of the information matrix.

Our Matlab implementation is based on the algorithm described on p. 312 in [14] and the code publicly available at [1]. We perform the Monte Carlo integration for various AR models with M integration points. But first, to test the accuracy we use the known result for the AR(1) model. Table 2 shows the fractional error obtained when $M = 10^5$ and $M = 10^6$. For models with larger order, we report the value $\Delta = |\hat{I}_{10^7} - \hat{I}_{10^6}|/\hat{I}_{10^7}$, where $\hat{I}_M$ denotes the Monte Carlo evaluation of $\int_\Theta |\mathbf{J}(g)|^{1/2} \mathrm{d}\theta$ calculated from M integration points. We show in Table 2 the results on Δ since it is known for Monte Carlo integration with Sobol' sequences that the fractional error decreases with the number of samples as $(\ln M)^n/M$ [14].

Example 2. We evaluate the capabilities of *NML*, *BIC* and *KICC* criteria for es-

TABLE 2
Monte Carlo results for the integral term in the stochastic complexity formula (4) for autoregressive models. For the AR(1) model the fractional error is reported.

M	$\hat{I}_M$	Fractional error or Δ
AR(1)		
10^5	3.131956	0.003067
10^6	3.138952	0.000840
AR(2) - pure complex poles		
10^6	42.06	-
10^7	47.41	0.11
AR(3) - one real-valued pole		
10^6	122.67	-
10^7	137.73	0.11
AR(4) - pure complex poles		
10^6	1069.66	-
10^7	1358.84	0.21
AR(5) - one real-valued pole		
10^6	3733.59	-
10^7	8307.55	0.55
AR(6) - pure complex poles		
10^6	23164.39	-
10^7	35981.48	0.36

timating the order of AR models. The *NML* criterion is calculated with formula (4), where the value of the integral term for $n > 1$ is the one from Table 2 computed with $M = 10^7$ integration points. We extend our experimental framework by considering another information theoretic criterion, namely the predictive least squares criterion *PLS*, [16].

Figure 1 outlines the simulation procedure used in Example 2, and the estimation results are shown in Tables 3-4.

Note that the evaluation of the various criteria for order estimation requires the

For the model order $n \in \{1, 2, 3\}$,
  For each order estimation criterion $\mathcal{C}$ and for each sample size N,
    $N \in \{25, 50, 100, 200\}$, initialize with zero two counters:
    $\mathcal{N}^c_{N,\mathcal{C}}$ for correct estimations and $\mathcal{N}^o_{N,\mathcal{C}}$ for over-estimations.
  Repeat the following steps 1000 times:
    Generate independently the entries of $\mathcal{P}_\mu$ as outcomes of $\mathcal{U}[(0.8, 1)]$,
      and the entries of $\mathcal{P}_\phi$ as outcomes of $\mathcal{U}[(0, \pi)]$.
      If n is odd, generate the unique entry of $\mathcal{P}_\rho$
      according to $\mathcal{U}\left[(-1, -0.8) \bigcup (0.8, 1)\right]$.
    Repeat the following steps 1000 times:
      Simulate a time series with 300 entries for the AR(n) process
        whose poles are given by $\mathcal{P}_\mu$, $\mathcal{P}_\phi$, $\mathcal{P}_\rho$.
        Use null initial conditions and $\sigma^2 = 1$.
        Discard the first 100 entries of the time series and
        dub z the vector formed with the rest of 200 measurements.
      For each sample size $N \in \{25, 50, 100, 200\}$,
          Choose $y^N = [z_1, \ldots, z_N]^\top$. Apply each criterion $\mathcal{C}$
          to estimate the model order $\hat{n}_{N,\mathcal{C}}$ from y^N data,
          under the hypothesis $\hat{n}_{N,\mathcal{C}} \in \{1, \ldots, 6\}$.
          If $\hat{n}_{N,\mathcal{C}} = n$, then increment $\mathcal{N}^c_{N,\mathcal{C}}$.
          If $\hat{n}_{N,\mathcal{C}} > n$, then increment $\mathcal{N}^o_{N,\mathcal{C}}$.
      End
    End
  End
  Calculate the probability of correct estimation $\hat{p}^c_{N,\mathcal{C}} = \mathcal{N}^c_{N,\mathcal{C}}/10^6$,
    and the probability of over-estimation $\hat{p}^o_{N,\mathcal{C}} = \mathcal{N}^o_{N,\mathcal{C}}/10^6$ for the model order.
End

FIG 1. *The simulation procedure applied in Example 2. The notation $\mathcal{U}[\cdot]$ is used for the uniform distribution.*

TABLE 3

Example 2 - the probability of correct estimation of the AR order. The best result for each sample size N is represented with bold font.

AR model order	Criterion	Sample size (N)			
		25	50	100	200
$n = 1$	*NML*	**0.99**	**0.99**	**1.00**	**1.00**
	BIC	0.93	0.95	0.97	0.98
	KICC	0.95	0.93	0.91	0.90
	PLS	0.89	0.92	0.95	0.97
$n = 2$	*NML*	0.72	**0.85**	**0.87**	**0.88**
	BIC	0.79	**0.85**	**0.87**	0.87
	KICC	**0.82**	0.83	0.80	0.78
	PLS	0.49	0.59	0.66	0.71
$n = 3$	*NML*	0.49	**0.74**	**0.83**	**0.84**
	BIC	**0.52**	0.71	0.78	0.79
	KICC	0.51	0.71	0.73	0.69
	PLS	0.26	0.39	0.47	0.53

TABLE 4
Example 2 - the probability to over-estimation of the order of AR models. The smallest overestimation probability for each sample size N is represented with bold font.

AR model order	Criterion	Sample size (N)			
		25	50	100	200
$n = 1$	*NML*	**0.01**	**0.01**	**0.00**	**0.00**
	BIC	0.07	0.05	0.03	0.02
	KICC	0.05	0.07	0.09	0.10
	PLS	0.11	0.08	0.05	0.03
$n = 2$	*NML*	0.07	**0.09**	**0.11**	**0.12**
	BIC	0.10	0.11	0.12	0.13
	KICC	**0.06**	0.14	0.20	0.22
	PLS	0.20	0.19	0.17	0.15
$n = 3$	*NML*	**0.01**	**0.03**	**0.06**	**0.12**
	BIC	0.07	0.09	0.12	0.18
	KICC	0.03	0.10	0.20	0.29
	PLS	0.21	0.22	0.23	0.23

estimate of noise variance for each order between one and six. Moreover, for the *PLS* criterion the computation of the prediction errors must be performed for each order and for each sample point. To reduce the computational burden, we resort to the fast implementation of the prewindowed estimation method based on predictive lattice filters [8], [22].

Observe in Table 3 that the *NML* criterion compares favorably with all the other criteria when the sample size is at least 50. For the smallest amount of data the asymptotic calculation of the Fisher information does not seem to be accurate enough. In most of the cases *BIC* is ranked the second after the *NML*, and the results of *KICC* do not improve when the sample size N is increased. For all criteria the performances decline for the larger values of the model order, which is clear because there is more to learn. Notice the moderate performances of the *PLS* criterion. We mention that another comparative study [7] also reports the moderate capabilities of *PLS* on estimating the order of AR models. This is to be expected since the *PLS* criterion is based on the estimates of the parameters which are shaky for small amounts of data.

5. ARMA models

The density function for ARMA models, (1), depends on how the initial values of y are related to the inputs e. A simple formula results if we put $y_i = e_i = 0$ for $i \leq 0$. Then the linear spaces spanned by y^t and e^t are the same. Let $\hat{y}_{t+1|t}$ be the orthogonal projection of y_{t+1} on the space spanned by y^t. We have the recursion

$$\hat{y}_{t+1|t} = \sum_{i=1}^{m} b_i (y_{t-i+1} - \hat{y}_{t-i+1|t-i}) - \sum_{i=1}^{n} a_i y_{t-i+1}, \tag{5}$$

where $\hat{y}_{1|0} = 0$. With more general initial conditions the coefficients b_i in (5) will depend on t; see for instance [17]. The likelihood function of the model is then

$$f(y^N; \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y_t - \hat{y}_{t|t-1})^2}. \tag{6}$$

TABLE 5

Results of model selection for the ARMA models in Example 3. The counts indicate for 1000 *runs the number of times the structure of the model was correctly estimated by each criterion, from the set* $\{ARMA(n,m) : n, m \geq 1, n+m \leq 6\}$. *The best result for each sample size N is represented with bold font.*

ARMA model	Criterion	Sample size (N)				
		25	50	100	200	400
$n=1, m=1$	*NML*	700	**812**	**917**	**962**	**989**
$a_1 = -0.5$	*BIC*	638	776	894	957	983
$b_1 = 0.8$	*KICC*	**717**	740	758	745	756
$n=2, m=1$	*NML*	**626**	**821**	**960**	**991**	**994**
$a_1 = 0.64, a_2 = 0.7$	*BIC*	532	740	898	961	978
$b_1 = 0.8$	*KICC*	586	727	810	846	849
$n=1, m=1$	*NML*	851	**887**	**918**	**931**	**961**
$a_1 = 0.3$	*BIC*	766	804	856	903	942
$b_1 = 0.5$	*KICC*	**860**	764	654	614	577

The maximized likelihood is $\dfrac{1}{(2\pi e\hat{\sigma}^2)^{N/2}}$, where $\hat{\sigma}^2 = \min\limits_{a_1,\ldots,a_n,b_1,\ldots,b_m} \dfrac{1}{N}\sum\limits_{t=1}^{N}(y_t - \hat{y}_{t|t-1})^2$. The *NML* criterion (3) is then given by

$$L(y^N; n, m) = \frac{N}{2}\ln(2\pi e\hat{\sigma}^2) + \frac{n+m+1}{2}\ln\frac{N}{2\pi} + \ln\int_{\Theta} |\mathbf{J}(\theta)|^{1/2}\mathrm{d}\theta + o(1), \tag{7}$$

In Appendix we elaborate on the computation of the integral term for the *NML* criterion, and the results are applied to the selection for ARMA models in the following example.

Example 3. We calculate the structure of ARMA models for data generated by three different processes, which also were used in [11]. For each model, the true structure and the coefficients are given in Table 5, where we show the estimation results for 1000 runs. In all experiments we have chosen the variance of the zero-mean white Gaussian noise to be $\sigma^2 = 1$. We mention that, similarly with the experiments on the autoregressive models each data set y^N was obtained after discarding the first 100 generated measurements. This is to eliminate the effect of the initial conditions. There exist different methods for estimation of ARMA models. We selected the one implemented in Matlab as armax function by Ljung, which is well described in his book [13].

Appendix: The asymptotic Fisher information matrix

We focus on the computation of the integral term in equation (7). The model is assumed to be stable and minimum phase, which means that in (2) the roots for both $B(q)$ and $A(q)$ are inside the open unit disc. Assume that n_1 zeros of $A(q)$ and m_1 zeros of $B(q)$ are real-valued. Then we have the inequalities $0 \leq n_1 \leq n$ and $0 \leq m_1 \leq m$. Because all coefficients of $A(q)$ and $B(q)$ are real-valued, the pure complex poles and zeros occur in complex conjugate pairs, and consequently the differences $n - n_1$ and $m - m_1$ are both even integers. For the pure complex poles and zeros we apply the parametrization in [5]:

$$g_{\ell+1} = g_\ell^* = |g_\ell|\exp(-i\phi_{g_\ell}), \quad \phi_{g_\ell} \in (0,\pi),\ \ell \in \{n_1+1, n_1+3, \ldots, n-1\},$$
$$h_{\ell+1} = h_\ell^* = |h_\ell|\exp(-i\phi_{h_\ell}), \quad \phi_{h_\ell} \in (0,\pi),\ \ell \in \{m_1+1, m_1+3, \ldots, m-1\},$$

where the symbol * denotes the complex conjugate. The entries of the parameter vector θ are given by:

$$\begin{aligned}\theta = (&g_1, \ldots, g_{n_1},\\ &|g_{n_1+1}|, \phi_{g_{n_1+1}}, \ldots, |g_{n-1}|, \phi_{g_{n-1}},\\ &h_1, \ldots, h_{m_1},\\ &|h_{m_1+1}|, \phi_{h_{m_1+1}}, \ldots, |h_{m-1}|, \phi_{h_{m-1}},\\ &\sigma^2).\end{aligned}$$

For the sake of clarity we define the subsets of indices for the θ parameters:

$$\begin{aligned}\mathcal{P}_\rho &= \{1, 2, \ldots, n_1\}\\ \mathcal{P}_\mu &= \{n_1+1, n_1+3, \ldots, n-1\}\\ \mathcal{P}_\phi &= \{n_1+2, n_1+4, \ldots n\}\\ \mathcal{P} &= \mathcal{P}_\rho \bigcup \mathcal{P}_\mu \bigcup \mathcal{P}_\phi\\ \mathcal{Z}_\rho &= \{n+1, n+2, \ldots n+m_1\}\\ \mathcal{Z}_\mu &= \{n+m_1+1, n+m_1+3, \ldots, n+m-1\}\\ \mathcal{Z}_\phi &= \{n+m_1+2, n+m_1+4, \ldots, n+m\}\\ \mathcal{Z} &= \mathcal{Z}_\rho \bigcup \mathcal{Z}_\mu \bigcup \mathcal{Z}_\phi\end{aligned}$$

Based on (6) we use the following asymptotic expression for the log-likelihood function of the observations $y_1, \ldots, y_N$, [2], [9]:

$$\mathcal{L} = -\frac{1}{2\sigma^2}\sum_{t=1}^{N} e_t^2 - \frac{N}{2}\ln\sigma^2 + \text{constant}.$$

For all $u, v \in \{1, \ldots, m+n+1\}$, the (u,v) entry of the Fisher information matrix is given by the formula [18]: $J_{u,v} = -\lim_{N\to\infty}\frac{1}{N}E[\frac{\partial^2\mathcal{L}}{\partial\theta_u\partial\theta_v}]$. Applying the results in [2] and [9], we obtain in a straightforward manner:

$$\begin{aligned}J_{n+m+1,n+m+1} &= 1/(2\sigma^4),\\ J_{u,n+m+1} = J_{n+m+1,v} &= 0 \ \forall u, v \in \{1, \ldots, n+m\}.\end{aligned}$$

For the following calculations we use the identity $J_{u,v} = \lim_{N\to\infty}\frac{1}{N}E[\frac{\partial\mathcal{L}}{\partial\theta_u}\frac{\partial\mathcal{L}}{\partial\theta_v}]$. Consider first the case $u, v \in \mathcal{P}_\rho$. Simple calculations lead to

$$\frac{\partial e_t}{\partial\theta_u} = -\frac{q^{-1}}{1-\theta_u q^{-1}}e_t = -\sum_{p=1}^{\infty}\theta_u^{p-1}q^{-p}e_t,$$

and we obtain readily:

$$\begin{aligned}J_{u,v} &= \frac{1}{N\sigma^4}E\left[\left(\sum_{t=1}^{N}e_t\sum_{p=1}^{\infty}\theta_u^{p-1}e_{t-p}\right)\left(\sum_{s=1}^{N}e_s\sum_{r=1}^{\infty}\theta_v^{r-1}e_{s-r}\right)\right]\\ &= \frac{1}{N\sigma^4}\sum_{t=1}^{N}\sum_{p=1}^{\infty}(\theta_u\theta_v)^{p-1}E\left[e_t^2e_{t-p}^2\right] \qquad (8)\\ &= \frac{1}{1-\theta_u\theta_v}.\end{aligned}$$

We conclude for $u, v \in \mathcal{P}_\rho \bigcup \mathcal{Z}_\rho$ that $J_{u,v} = \dfrac{\mathcal{S}_u \mathcal{S}_v}{1 - \theta_u \theta_v}$, where

$$\mathcal{S}_u = \begin{cases} -1, u \in \mathcal{P} \\ 1, \ u \in \mathcal{Z} \end{cases}$$

Formula (8) was deduced in [4] for the case when all the poles and the zeros of the ARMA(n,m) model are real-valued. We evaluate next the entry (u, v) of the Fisher information matrix for $u \in \mathcal{P}_\rho \bigcup \mathcal{Z}_\rho$ and $v \in \mathcal{P}_\mu \bigcup \mathcal{P}_\phi \bigcup \mathcal{Z}_\mu \bigcup \mathcal{Z}_\phi$. It is not difficult to prove that

$$\frac{\partial e_s}{\partial \theta_v} = \sum_{r=1}^{\infty} d_{v,r} e_{s-r} \ \forall s \in \{1, \ldots, N\},$$

where the coefficients $d_{v,r}$ are real-valued, [5]. Therefore

$$\begin{aligned} J_{u,v} &= \frac{\mathcal{S}_u}{N\sigma^4} E\left[\left(\sum_{t=1}^{N} e_t \sum_{p=1}^{\infty} \theta_u^{p-1} e_{t-p}\right)\left(\sum_{s=1}^{N} e_s \sum_{r=1}^{\infty} d_{v,r} e_{s-r}\right)\right] \\ &= \frac{\mathcal{S}_u}{N\sigma^4} \sum_{t=1}^{N} \sum_{p=1}^{\infty} \theta_u^{p-1} d_{v,p} E\left[e_t^2 e_{t-p}^2\right] \\ &= \mathcal{S}_u \sum_{p=1}^{\infty} \theta_u^{p-1} d_{v,p}. \end{aligned}$$

The following closed form expressions of $d_{v,p}$ are given in [5] for $v \in \mathcal{P}_\mu \bigcup \mathcal{Z}_\mu$:

$$d_{v,p} = \begin{cases} 2\mathcal{S}_v \cos\theta_{v+1}, & p = 1 \\ 2\mathcal{S}_v \frac{\theta_v^p \sin(p\theta_{v+1}) \cos\theta_{v+1} - \theta_v^{p-1} \sin((p-1)\theta_{v+1})\theta_v}{\theta_v \sin\theta_{v+1}}, & p \geq 2 \end{cases}$$

The equations above lead to

$$\begin{aligned} J_{u,v} &= 2\frac{\mathcal{S}_u \mathcal{S}_v}{\theta_u \theta_v} \frac{\cos\theta_{v+1}}{\sin\theta_{v+1}} \sum_{p=1}^{\infty} (\theta_u \theta_v)^p \sin(p\theta_{v+1}) \\ &\quad -2\frac{\mathcal{S}_u \mathcal{S}_v}{\sin\theta_{v+1}} \sum_{p=1}^{\infty} (\theta_u \theta_v)^p \sin(p\theta_{v+1}) \\ &= 2\mathcal{S}_u \mathcal{S}_v \frac{\cos\theta_{v+1} - \theta_u \theta_v}{1 - 2\theta_u \theta_v \cos\theta_{v+1} + \theta_u^2 \theta_v^2}, \end{aligned}$$

for $u \in \mathcal{P}_\rho$ and $v \in \mathcal{P}_\mu \bigcup \mathcal{Z}_\mu$. Similarly for $v \in \mathcal{P}_\phi \bigcup \mathcal{Z}_\phi$ and $p \geq 1$, we have, [5],

$$d_{v,p} = -2\mathcal{S}_v \theta_{v-1}^p \sin(p\theta_v),$$

and it is easy to prove that

$$J_{u,v} = -2\mathcal{S}_u \mathcal{S}_v \frac{\theta_{v-1} \sin\theta_v}{1 - 2\theta_u \theta_{v-1} \cos\theta_v + \theta_u^2 \theta_{v-1}^2}.$$

When $u, v \in \mathcal{P}_\mu \bigcup \mathcal{Z}_\mu \bigcup \mathcal{P}_\phi \bigcup \mathcal{Z}_\phi$, we can apply the formulas given in [5] for the computation of $J_{u,v}$ in case all the poles and the zeros are purely complex.

Analyzing the sign of the product $\mathcal{S}_u\mathcal{S}_v$, we find that the matrix $\mathbf{J}(\theta)$ can be re-written more compactly as $\mathbf{J}(\theta) = \begin{bmatrix} \mathbf{G} & -\mathbf{C} \\ -\mathbf{C}^\top & \mathbf{H} \end{bmatrix}$, where the size of the block matrix $\mathbf{C}$ is $n \times m$. The identity $\begin{vmatrix} \mathbf{G} & -\mathbf{C} \\ -\mathbf{C}^\top & \mathbf{H} \end{vmatrix} = \begin{vmatrix} \mathbf{G} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{H} \end{vmatrix}$ leads to the conclusion that $\int_\Theta |\mathbf{J}(\theta)|^{1/2} \mathrm{d}\theta$ has the same value for the models ARMA(n,m), ARMA(n+m,0), ARMA(0,n+m). A similar conclusion was drawn in [4] for the particular case when all the poles and the zeros are real-valued.

References

[1] http://www2.math.uic.edu/∼hanson/mcs507/cp4f04.html.

[2] Åström, K. (1967). On the achievable accuracy in identification problems. In *Preprints of the IFAC Symposium Identification in Automatic Control Systems.* Prague, Czechoslovakia, 9 pp.

[3] Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **44** 2743–2760.

[4] Box, G. and Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control.* Holden-Day, Inc.

[5] Bruzzone, S. and Kaveh, M. (1984). Information tradeoffs in using the sample autocorrelation function in ARMA parameter estimation. *IEEE Trans. on Acoustics, Speech and Signal Processing* **ASSP-32** (4, Aug.) 701–715.

[6] Cavanaugh, J. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statist. Probability Lett.* **42** 333–343.

[7] Djuric, P. and Kay, S. (1992). Order selection of autoregressive models. *IEEE Trans. Signal. Proces.* **40** 2829–2833.

[8] Friedlander, B. (1982). Lattice filters for adaptive processing. *Proc. IEEE* **70** 829–868.

[9] Friedlander, B. (1984). On the computation of the Cramer-Rao bound for ARMA parameter estimation. *IEEE Trans. on Acoustics, Speech and Signal Processing* **ASSP-32** (4, Aug.) 721–727.

[10] Friedlander, B. and Porat, B. (1989). The exact Cramer-Rao bound for Gaussian autoregressive processes. *IEEE Tr. on Aerospace and Electronic Systems* **AES-25** 3–8.

[11] Hannan, E. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69** (1) 81–94.

[12] Kay, S. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice-Hall, Inc.

[13] Ljung, L. (1999). *System Identification: Theory for the User*, 2nd ed. Prentice Hall, Upper Saddle River, NJ.

[14] Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (1992). *Numerical Recipies in C. The Art of Scientific Computing*, 2nd ed. Cambridge University Press.

[15] Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

[16] Rissanen, J. (1986). Order estimation by acumulated prediction errors. *J. Appl. Prob.* **23A** 55–61.

[17] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific Publ. Co., River Edge, NJ, 175 pp.

[18] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **42** (1, Jan.) 40–47.

[19] Rissanen, J. (2000). MDL denoising. *IEEE Trans. Inf. Theory* **46** (7, Nov.) 2537–2543.

[20] Schwarz, G. (1978). Estimating the dimension of the model. *Ann. Stat.* **6** 461–464.

[21] Seghouane, A.-K. and Bekara, M. (2004). A small sample model selection criterion based on Kullback's symmetric divergence. *IEEE Trans. Signal. Proces.* **52** 3314–3323.

[22] Wax, M. (1988). Order selection for AR models by predictive least squares. *IEEE Trans. on Acoustics, Speech and Signal Processing* **36** 581–588.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 60–71

DOI: 10.1214/074921706000000950

On prediction errors in regression models with nonstationary regressors

Ching-Kang Ing[1] and Chor-Yiu Sin[2]

Academia Sinica and Xiamen University

Abstract: In this article asymptotic expressions for the final prediction error (FPE) and the accumulated prediction error (APE) of the least squares predictor are obtained in regression models with nonstationary regressors. It is shown that the term of order $1/n$ in FPE and the term of order $\log n$ in APE share the same constant, where n is the sample size. Since the model includes the random walk model as a special case, these asymptotic expressions extend some of the results in Wei (1987) and Ing (2001). In addition, we also show that while the FPE of the least squares predictor is not affected by the contemporary correlation between the innovations in input and output variables, the mean squared error of the least squares estimate does vary with this correlation.

1. Introduction

Consider a simple regression model

$$y_t = \beta x_{t-1} + \varepsilon_t, \tag{1.1}$$

where β is an unknown constant, ε_t's are (unobservable) independent random disturbances with zero means and a common variance σ^2, and x_t is an unit root process satisfying

$$x_t = x_{t-1} + \eta_t, \tag{1.2}$$

with $x_0 = 0$, $\eta_t = \sum_{j=0}^{t-1} c_j \omega_{t-j}$, $\sum_{j=0}^{\infty} |c_j| < \infty$, $\sum_{j=0}^{\infty} c_j \neq 0$, and ω_t being independent random noises with zero means and a common variance σ_ω^2. We also assume that ε_t is independent of $\{\omega_j, j \leq t-1\}$. Note that if $\beta = 1$, $c_0 = 1$, $c_j = 0$ if $j > 0$, and $\varepsilon_t = \omega_t$, then (1.1) becomes the well-known random walk model (see, for instance, Chan and Wei [4]). Having observed $(y_{i+1}, x_i), i = 1, \ldots, n-1$, β can be estimated by least squares

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n-1} x_i y_{i+1}}{\sum_{i=1}^{n-1} x_i^2}. \tag{1.3}$$

If x_n also becomes available, then it is natural to predict y_{n+1} using the least squares predictor,

$$\hat{y}_{n+1} = x_n \hat{\beta}_n. \tag{1.4}$$

[1]Academia Sinica, Taipei, Taiwan, R.O.C., e-mail: cking@stat.sinica.edu.tw

[2]Wang Yanan Institute for Studies in Economics, Xiamen University, Fujian, e-mail: cysinhkbu@gmail.com

AMS 2000 subject classifications: primary 60M20; secondary 62F12, 62M10.

Keywords and phrases: accumulated prediction errors, final prediction error, least squares estimators, random walk models.

To assess the performances of the least squares predictor, we consider the final prediction error (FPE, Akaike [1])

$$E\left\{ (\, y_{n+1} - \hat{y}_{n+1}\,)^2 \right\} = \sigma^2 + E\left\{ x_n^2 (\, \hat{\beta}_n - \beta\,)^2 \right\}, \tag{1.5}$$

and the accumulated prediction error (APE, Rissanen [14])

$$\begin{aligned} \sum_{i=2}^{n} (y_i - \hat{y}_i)^2 &= \sum_{i=2}^{n} \left\{ \varepsilon_i - x_{i-1}(\hat{\beta}_{i-1} - \beta) \right\}^2 \\ &= \sum_{i=2}^{n} \varepsilon_i^2 + \sum_{i=2}^{n} x_{i-1}^2 (\hat{\beta}_{i-1} - \beta)^2 (1 + o(1)) \text{ a.s.}, \end{aligned} \tag{1.6}$$

where the second equality of (1.6) is ensured by Chow [5]. It is straightforward to see that the terms in (1.5) and (1.6),

$$\sum_{i=2}^{n} x_{i-1}^2 (\hat{\beta}_{i-1} - \beta)^2 = \sum_{i=2}^{n} \left\{ \frac{x_i^2 (\sum_{j=1}^{i-1} x_j \varepsilon_{j+1})^2}{(\sum_{j=1}^{i-1} x_j^2)^2} \right\}, \tag{1.7}$$

and

$$n x_n^2 (\hat{\beta}_n - \beta\,)^2 = \left\{ \frac{(\frac{1}{\sqrt{n}} x_n)(\frac{1}{n} \sum_{i=1}^{n-1} x_i \varepsilon_{i+1})}{\frac{1}{n^2} \sum_{i=1}^{n-1} x_i^2} \right\}^2. \tag{1.8}$$

When $\{y_t\}$ is a random walk model mentioned above, Wei ([15], Theorem 4) showed that the rhs of (1.7) equals $2\sigma_\omega^2 \log n + o(\log n)$ a.s. By imposing further assumptions on the distribution of ω_t, Ing ([9], Corollary 1) subsequently obtained the limiting value of the expectation on the rhs of (1.8), which is $2\sigma_\omega^2$. This article extends these two results to models (1.1) and (1.2), which provides a deeper understanding of the least squares predictor (estimate) in situations where Fisher's information, $\sum_{j=1}^{n-1} x_j^2$, grows at a rate much faster than n, and the innovations in input and output variables come from different sources. The rest of the paper is organized as follows. Section 2 derives the asymptotic expressions for the rhs of (1.7). In Section 3, sufficient conditions are given to ensure that the expectation on the rhs of (1.8) is bounded by some finite positive constant for all sufficiently large n. We then apply this moment property and the results obtained in Section 2 to show that

$$\lim_{n \to \infty} E\{ n x_n^2 (\hat{\beta}_n - \beta\,)^2 \} = 2\sigma^2. \tag{1.9}$$

Some discussions related to (1.9) are given at the end of Section 3. In particular, it is shown that while the FPE of the least squares predictor is not affected by the contemporary correlation between ε_t and ω_t, the mean squared error of the least squares estimate does vary with this correlation. In addition, we also show that the squares of the normalized estimate, $n(\hat{\beta}_n - \beta)$, and the normalized regressor, $x_n/\sqrt{n}$, are not asymptotically uncorrelated.

2. An asymptotic expression for the APE

To prove the main result of this section, two auxiliary lemmas are required. They are also of independent interests.

Lemma 1. *Assume the* $\{\omega_t\}$ *in Section 1 satisfy* $\sup_{-\infty<t<\infty} E|\omega_t|^\alpha < \infty$ *for some* $\alpha > 2$. *Let* $z_t = \sum_{j=0}^{t-1} d_j \omega_{t-j}$, *where* $|d_j| \le Cj^{-1}$ *for some* $C > 0$ *and all* $j \ge 1$. *Then, with* $\gamma_t = \sigma_\omega^2 \sum_{j=0}^{t-1} d_j^2$,

$$\frac{1}{n}\sum_{t=1}^{n}(z_t^2 - \gamma_t) = o(1) \text{ a.s.} \tag{2.1}$$

Proof. Straightforward calculations yield that

$$z_t^2 - \gamma_t = \sum_{l=1}^{t} d_{t-l}^2(\omega_l^2 - \sigma_\omega^2) + 2\sum_{l_2=2}^{t}\sum_{l_1=1}^{l_2-1} d_{t-l_1} d_{t-l_2}\omega_{l_1}\omega_{l_2}. \tag{2.2}$$

By (2.2) and changing the order of summations,

$$\begin{aligned}\sum_{t=n_1}^{n_2}\frac{z_t^2-\gamma_t}{t} &= \sum_{l=1}^{n_1}\left(\sum_{t=n_1}^{n_2}\frac{d_{t-l}^2}{t}\right)\eta_l^* + \sum_{l=n_1+1}^{n_2}\left(\sum_{t=l}^{n_2}\frac{d_{t-l}^2}{t}\right)\eta_l^* \\ &\quad + 2\sum_{l_2=2}^{n_1}\left\{\sum_{l_1=1}^{l_2-1}\left(\sum_{t=n_1}^{n_2}\frac{d_{t-l_1}d_{t-l_2}}{t}\right)\omega_{l_1}\right\}\omega_{l_2} \\ &\quad + 2\sum_{l_2=n_1+1}^{n_2}\left\{\sum_{l_1=1}^{l_2-1}\left(\sum_{t=l_2}^{n_2}\frac{d_{t-l_1}d_{t-l_2}}{t}\right)\omega_{l_1}\right\}\omega_{l_2} \\ &\equiv (1)+(2)+(3)+(4),\end{aligned}$$

where $\eta_t^* = \omega_t^2 - \sigma_\omega^2$. In the following, we shall show that for some $\alpha_k > 1$, there are $C_k > 0, \xi_{1,k} > 1$, and $\xi_{2,k} > 1$ independent of n_1 and n_2 such that

$$E|(k)|^{\alpha_k} \le C_k\Big(\sum_{t=n_1}^{n_2}\frac{1}{t^{\xi_{1,k}}}\Big)^{\xi_{2,k}}, \tag{2.3}$$

where $k = 1, \ldots, 4$. (2.3) and Móricz (1976) imply that for some $\alpha > 1$, there are $C^* > 0, \xi_1 > 1$, and $\xi_2 > 1$ independent of n_1 and n_2 such that

$$E\max_{n_1\le l\le n_2}\Big|\sum_{t=n_1}^{l}\frac{z_t^2-\gamma_t}{t}\Big|^\alpha \le C^*\Big(\sum_{t=n_1}^{n_2}\frac{1}{t^{\xi_1}}\Big)^{\xi_2}. \tag{2.4}$$

As a result, (2.1) follows from (2.4) and Kronecker's lemma.

Let $\alpha_1 = \min\{\alpha/2, 2\}$. Then,

$$\begin{aligned}E|(1)|^{\alpha_1} &\le C_{1,1}E\Big\{\sum_{l=1}^{n_1}\Big(\sum_{t=n_1}^{n_2}\frac{d_{t-l}^2}{t}\Big)^2 {\eta_l^*}^2\Big\}^{\alpha_1/2} \\ &\le C_{1,1}\sum_{t_1=n_1}^{n_2}\sum_{t_2=n_1}^{n_2}\frac{1}{t_1^{\alpha_1/2}t_2^{\alpha_1/2}}\sum_{l=1}^{n_1}|d_{t_1-l}d_{t_2-l}|^{\alpha_1}E|\eta_l^*|^{\alpha_1} \\ &\le C_{1,2}\left(\sum_{t=n_1}^{n_2}\frac{1}{t^{\alpha_1}} + \sum_{t_1=n_1}^{n_2-1}\frac{1}{t_1^{\alpha_1/2}}\sum_{t_2=t_1+1}^{n_2}\frac{1}{t_2^{\alpha_1/2}}(t_2-t_1)^{-\alpha_1}\right) \\ &\le C_{1,3}\left(\sum_{t=n_1}^{n_2}\frac{1}{t^{\alpha_1}}\right) \le C_{1,3}\left(\sum_{t=n_1}^{n_2}\frac{1}{t^{\xi_{1,1}}}\right)^{\xi_{2,1}},\end{aligned} \tag{2.5}$$

where $C_{1,i}, i = 1, 2, 3$ are some positive constant independent of n_1 and n_2, $1 < \xi_{1,1} < \alpha_1$, $\xi_{2,1} = \alpha_1/\xi_{1,1}$, first inequality follows from Burkholder's inequality, second one follows from the fact that $0 < \alpha_1/2 \leq 1$ and changing the order of summations, third one is ensured by $\sup_t E|\omega_t|^\alpha < \infty$ and $|d_j| \leq Cj^{-1}$, which implies for all $n_1 \leq t_1, t_2 \leq n_2$, $\sum_{l=1}^{n_1} |d_{t_1-l} d_{t_2-l}|^{\alpha_1} \leq C_{1,4}|t_1 - t_2|^{-\alpha_1}$, for some $C_{1,4} > 0$. As a result, (2.3) holds for $k = 1$. The proof of (2.3) for the case of $k = 2$ is similar. The details are thus omitted. To show (2.3) for the case $k = 3$, let $\alpha_3 = \alpha$. Then, by Minkowski's inequality and using Wei (1987, Lemma 2) twice, one obtains

$$\begin{aligned} E|(3)|^{\alpha_3} &\leq C_{3,1} E| \sum_{l_2=2}^{n_1} \left\{ \sum_{l_1=1}^{l_2-1} \left(\sum_{t=n_1}^{n_2} \frac{d_{t-l_1} d_{t-l_2}}{t} \right) \omega_{l_1} \right\} \omega_{l_2}|^{\alpha_3} \\ &\leq C_{3,2} \left(\sum_{l_2=2}^{n_1} \sum_{l_1=1}^{l_2-1} \Big(\sum_{t=n_1}^{n_2} \frac{d_{t-l_1} d_{t-l_2}}{t} \Big)^2 \right)^{\alpha_3/2}, \end{aligned} \tag{2.6}$$

where $C_{3,i}, i = 1, 2$ are some positive constants independent of n_1 and n_2. Observe that for $n_1 \leq t_1 < t_2 \leq n_2$ and any $1 \leq M_1 \leq M_2 \leq n_1$, $\sum_{l=M_1}^{M_2} |d_{t_1-l} d_{t_2-l}| \leq C_{3,3}(\log t_2 - \log t_1)/(t_2 - t_1)$, where $C_{3,3} > 0$ is independent of M_1 and M_2. Using this fact and changing the order of summations, it follows that the rhs of (2.6) is bounded by $C_{3,4}(\sum_{t=n_1}^{n_2} t^{-2})^{\alpha_3/2}$, where $C_{3,4}$ is a positive constant independent of n_1 and n_2. Hence, (2.3) holds for $k = 3$. The proof of (2.3) for the case $k = 4$ is similar to that of $k = 3$. Therefore, we skip the details. □

Remark 1. If in Lemma 1 $z_t = \sum_{j=0}^{\infty} d_j \omega_{t-j}$ with $|d_j| \leq Cj^{-1}, j \geq 1$, then the same argument also yields (2.1) but with γ_t replaced by $\gamma^* = \sigma_\omega^2 \sum_{j=0}^{\infty} d_j^2$. For a related result, Brockwell and Davis (1987, Proposition 7.3.5), assuming that ω_j's are i.i.d. with finite second moment and d_j's satisfy $\sum_{j=0}^{\infty} |d_j| < \infty$ and $\sum_{j=0}^{\infty} d_j^2 j < \infty$, obtained $(n^{-1} \sum_{t=1}^{n} z_t^2) - \gamma^* = o_p(1)$. While the moment restriction of their result is slightly weaker than that of Lemma 1, the identically distributed assumption can be dropped in Lemma 1. In addition, the assumption on d_j in Lemma 1 seems less stringent. More importantly, Lemma 1 gives a *strong law* of large number for $n^{-1} \sum_{t=1}^{n} z_t^2$ under rather mild assumptions, which is one of the key tools for our asymptotic analysis of APE.

Lemma 2. *Assume* $\sup_{-\infty<t<\infty} E|\omega_t|^\alpha < \infty$ *for some* $\alpha > 2$ *and*

$$\sum_{j \geq k} |c_j| = O(k^{-1}). \tag{2.7}$$

Then,

$$\log \left(\sum_{j=1}^{n-1} x_j^2 \right) = 2 \log n + o(\log n) \text{ a.s.}$$

Proof. First note that $x_t = \sum_{j=1}^{t} \eta_j$. Define $N_t = \theta \sum_{j=1}^{t} \omega_j$, where $\theta = \sum_{j=0}^{\infty} c_j$. Then,

$$x_t = N_t - S_t, \tag{2.8}$$

where $S_t = \sum_{j=0}^{t-1} f_j \omega_{t-j}$ with $f_j = \sum_{l=j+1}^{\infty} c_l$. In view of (2.8),

$$\sum_{j=1}^{n-1} x_j^2 = \sum_{j=1}^{n-1} N_j^2 - 2 \sum_{j=1}^{n-1} N_j S_j + \sum_{j=1}^{n-1} S_j^2. \tag{2.9}$$

Since $|f_j| = O(j^{-1})$, Lemma 1 yields

$$\sum_{j=1}^{n-1} S_j^2 = O(n) \text{ a.s.} \tag{2.10}$$

By the law of the iterated logarithm,

$$\sum_{j=1}^{n-1} N_j^2 = O(n^2 \log\log n) \text{ a.s.} \tag{2.11}$$

By Lai and Wei ([12], (3.23)),

$$\liminf_{n\to\infty} \frac{\log\log n}{n^2} \sum_{j=1}^{n-1} N_j^2 > 0 \text{ a.s.} \tag{2.12}$$

Now, Lemma 2 follows directly from (2.9)-(2.12). □

Remark 2. By assuming

$$\sum_{j=0}^{\infty} j|c_j| < \infty, \tag{2.13}$$

Proposition 17.3 of Hamilton (1994) gives the limiting distribution of $n^{-2}\sum_{j=1}^{n-1} x_j^2$, which is $\lambda^2 \int_0^1 w(r)^2 dr$, where $\lambda = \sigma_\omega \sum_{j=0}^{\infty} c_j$ and $w(r)$ denotes the standard Brownian motion. This result immediately implies

$$\log\left(\sum_{j=1}^{n-1} x_j^2\right) = 2\log n + O_p(1). \tag{2.14}$$

Lemma 2 and (2.14) provide different estimates for the difference between $2\log n$ and $\log(\sum_{j=1}^{n-1} x_j^2)$, but neither is more informative than the other. On the other hand, we have found that the assumption on the coefficients used in Lemma 2, (2.7), seems to be weaker than the one imposed by Hamilton, (2.13). This can be seen by observing that (2.7) is marginally satisfied by $C_1 j^{-2} \le |c_j| \le C_2 j^{-2}, C_2 \ge C_1 > 0$, whereas (2.13) is not.

We are now ready to prove the main result of this section.

Theorem 1. *Assume that models (1.1), (1.2), and the assumptions of Lemma 2 hold. Also assume that* $\sup_{-\infty<t<\infty} E|\varepsilon_t|^{\alpha_0} < \infty$ *for some* $\alpha_0 > 2$. *Then,*

$$\sum_{i=2}^{n} x_{i-1}^2 (\hat{\beta}_{i-1} - \beta)^2 = 2\sigma^2 \log n + o(\log n) \text{ a.s.}, \tag{2.15}$$

and

$$\sum_{i=2}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=2}^{n} \varepsilon_i^2 + 2\sigma^2 \log n + o(\log n) \text{ a.s.} \tag{2.16}$$

Proof. First note that (2.9)-(2.12) yield

$$\limsup_{n\to\infty} \frac{n^2}{(\log\log n)\sum_{j=1}^{n} x_j^2} < \infty \text{ a.s.} \tag{2.17}$$

By Wei ([15], Lemma 2) and (2.7),

$$E\left|\frac{S_n}{n^{1/2}}\right|^{\alpha} \le C_\alpha n^{-\alpha/2}\left(\sum_{j=0}^{n-1} f_j^2\right)^{\alpha} \le C_\alpha^* n^{-\alpha/2}, \tag{2.18}$$

where C_α and C_α^* depend only on α. (2.18) and the Borel-Cantelli lemma give

$$S_n = o(n^{1/2}) \text{ a.s.} \tag{2.19}$$

Since the law of the iterated logarithm implies

$$N_n = O((n\log\log n)^{1/2}) \text{ a.s.},$$

this, (2.8), (2.17), and (2.19) yield

$$\frac{x_n^2}{\sum_{j=1}^{n} x_j^2} = o(1) \text{ a.s.} \tag{2.20}$$

In view of (2.20) and Wei ([15], Theorem 3), we have

$$\sum_{i=2}^{n} x_{i-1}^2(\hat\beta_{i-1}-\beta)^2 = \sigma^2 \log\left(\sum_{j=1}^{n-1} x_j^2\right) + o\left(\log\left(\sum_{j=1}^{n-1} x_j^2\right)\right) \text{ a.s.}, \tag{2.21}$$

As a result, (2.15) follows from Lemma 2 and (2.21); and (2.16) is an immediate consequence of (2.15) and (1.6). □

3. An asymptotic expression for the FPE

Assume that models (1.1) and (1.2) hold, $E(\varepsilon_t\omega_t) = \pi$ is a constant independent of t, $\sup_{-\infty<t<\infty} E|\varepsilon_t|^{\alpha_0} < \infty, \alpha_0 > 2$, and $\sup_{-\infty<t<\infty} E|\omega_t|^{\alpha} < \infty, \alpha > 2$. Then, by the functional central limit theorem, continuous mapping theorem, Ito's formula, and some algebraic manipulations, it can be shown that

$$\begin{aligned} &\left\{\frac{(\frac{1}{\sqrt{n}}x_n)(\frac{1}{n}\sum_{i=1}^{n-1} x_i\varepsilon_{i+1})}{\frac{1}{n^2}\sum_{i=1}^{n-1} x_i^2}\right\}^2 \\ &\quad\Longrightarrow \frac{w_a^2(1)\left(\rho\sigma_\omega\int_0^1 w_a(t)dw_a(t) + \sigma_\theta\int_0^1 w_a(t)dw_b(t)\right)^2}{\left(\int_0^1 w_a^2(t)dt\right)^2}, \end{aligned} \tag{3.1}$$

where "$\Longrightarrow$" denotes weak convergence, $(w_a(t), w_b(t))$ is a standard Brownian motion of dimension 2, $\rho = \pi/\sigma_\omega^2$, and $\sigma_\theta^2 = \sigma^2 - \rho^2\sigma_\omega^2$. If we can further show that for some $q > 2$,

$$E\left|\frac{(\frac{1}{\sqrt{n}}x_n)(\frac{1}{n}\sum_{i=1}^{n-1} x_i\varepsilon_{i+1})}{\frac{1}{n^2}\sum_{i=1}^{n-1} x_i^2}\right|^q = O(1), \tag{3.2}$$

then, in view of (3.1), (3.2), and (1.8),

$$nE\{x_n^2(\hat{\beta}_n-\beta)^2\} = E\left\{\frac{w_a^2(1)\left(\rho\sigma_\omega\int_0^1 w_a(t)dw_a(t)+\sigma_\theta\int_0^1 w_a(t)dw_b(t)\right)^2}{\left(\int_0^1 w_a^2(t)dt\right)^2}\right\}+o(1). \tag{3.3}$$

In the rest of this section, we provide sufficient conditions to ensure (3.2). In addition, the expectation on the rhs of (3.3) is investigated (Corollary 1). Let us start with a useful lemma.

Lemma 3. *Let $F_{t,m,\mathbf{a}_m}(.)$ be the distribution function of $\sum_{j=1}^m a_j\omega_{t+1-j}$, where $\mathbf{a}_m=(a_1,\ldots,a_m)'$. There are some positive numbers κ, ι, and M such that for all $m\geq 1$,$-\infty<t<\infty$ and $\|\mathbf{a}_m\|^2=\sum_{j=1}^m a_j^2=1$,*

$$\mid F_{t,m,\mathbf{a}_m}(x)-F_{t,m,\mathbf{a}_m}(y)\mid\leq M\mid x-y\mid^\kappa, \tag{3.4}$$

as $\mid x-y\mid\leq\iota$. Then, for any $q>0$,

$$E\left\{\left(\frac{1}{n^2}\sum_{j=1}^{n-1}x_j^2\right)^{-q}\right\}=O(1). \tag{3.5}$$

Proof. The proof is closely related to the one given in Ing ([9], Lemma 1), with the assumption there being strengthened to (3.4). First note that

$$\frac{1}{n^2}\sum_{i=1}^{n-1}x_i^2\geq\frac{1}{n^2}\sum_{i=n\delta}^{n-1}x_i^2=\frac{\delta}{n}\sum_{i=n\delta}^{n-1}\frac{x_i^2}{n\delta}\geq\frac{\delta}{n}\sum_{i=n\delta}^{n-1}\frac{x_i^2}{i}, \tag{3.6}$$

where $0<\delta<1$, and without loss of generality, $n\delta$ is assumed to be a positive integer. Rearranging the series on the rhs of (3.6), one obtains

$$\frac{\delta}{n}\sum_{j=0}^{\frac{(1-\delta)n}{lq}-1}\sum_{i=0}^{lq-1}\frac{x^2_{n\delta+\frac{(1-\delta)n}{lq}i+j}}{n\delta+\frac{(1-\delta)n}{lq}i+j}, \tag{3.7}$$

where $l>\max[\,2/\kappa,1/q,(1/q)\{(1/\delta)-1)\}\,]$ and for simplifying the discussion, lq and $\{\,(1-\delta)n\,\}/(lq)$ are also assumed to be positive integers. By the convexity of function $x^{-q},x>0$,

$$\left(\frac{1}{n^2}\sum_{i=1}^{n-1}x_i^2\right)^{-q}\leq\left\{\frac{(1-\delta)\,\delta}{lq}\right\}^{-q}\frac{lq}{(1-\delta)\,n}\times\sum_{j=0}^{\frac{(1-\delta)n}{lq}-1}\left\{\sum_{i=0}^{lq-1}\frac{x^2_{n\delta+\frac{(1-\delta)n}{lq}i+j}}{n\delta+\frac{(1-\delta)n}{lq}i+j}\right\}^{-q}. \tag{3.8}$$

In view of (3.8), if one can show that for some positive number C independent of j, the following inequality,

$$E\left\{\sum_{i=0}^{lq-1}\frac{x^2_{n\delta+\frac{(1-\delta)n}{lq}i+j}}{n\delta+\frac{(1-\delta)n}{lq}i+j}\right\}^{-q}\leq C<\infty, \tag{3.9}$$

holds for all $j = 0, 1, \ldots, \{(1-\delta)n/(lq)\} - 1$ as n is large enough, then (3.5) follows. The rest of the proof only focuses on the case where $j = 0$, because the same argument can be easily applied to other j's.

For $i = 0, \ldots, lq-1$, define

$$Y_{n,i} = \left\{ n\delta + \frac{(1-\delta)\,n}{lq} i \right\}^{-1/2} x_{n\delta + \frac{(1-\delta)n}{lq} i}, \tag{3.10}$$

$$W_{n,i} = \left\{ n\delta + \frac{(1-\delta)\,n}{lq} i \right\}^{-1/2} \sum_{m=0}^{\frac{(1-\delta)n}{lq}-1} \bar{f}_m \omega_{n\delta + \frac{(1-\delta)n}{lq} i - m}, \tag{3.11}$$

where $\bar{f}_j = \sum_{l=0}^{j} c_l$, and

$$F_{n,i} = Y_{n,i} - W_{n,i}. \tag{3.12}$$

(Note that $x_t = \sum_{j=0}^{t-1} \bar{f}_j \omega_{t-j}$.) Then,

$$\begin{aligned} E\left(\sum_{i=0}^{lq-1} Y_{n,i}^2\right)^{-q} &= \int_0^\infty Pr\left\{ \left(\sum_{i=0}^{lq-1} Y_{n,i}^2\right)^{-q} > t \right\} dt \\ &= \int_0^\infty Pr\left(\sum_{i=0}^{lq-1} Y_{n,i}^2 < t^{-1/q}\right) dt \\ &\le \int_0^\infty Pr\left(-t^{-1/(2q)} < Y_{n,i} < t^{-1/(2q)}, \quad i = 0, \ldots, lq-1\right) dt \\ &= \int_0^\infty E\left\{ E\left(\prod_{i=0}^{lq-1} I_{A_{n,i}} \,\middle|\, F_{n,lq-1}, W_{n,i}, F_{n,i},\ i = 0, \ldots, lq-2 \right) \right\} dt, \end{aligned} \tag{3.13}$$

where $A_{n,i} = \{ -t^{-1/(2q)} < Y_{n,i} < t^{-1/(2q)} \}$. In view of (3.10)-(3.12), for $0 \le p \le lq-1$, $0 \le i \le p$, and $0 \le j \le p-1$, $W_{n,p}$ is independent of $(F_{n,i}, W_{n,j})$. In addition, $\mathrm{var}(W_{n,i}) > \zeta > 0$, where $i = 0, \ldots, lq-1$ and ζ is a positive number independent of n and i. According to these facts, (3.4), and arguments similar to those used in (3.10) and (3.11) of Ing [9], there exist some positive numbers $0 < C' < \infty, 0 < s < \infty$, and a positive integer N_0 such that for all $n \ge N_0$ and all $t \ge s$,

$$E\left(\prod_{i=0}^{lq-1} I_{A_{n,i}} \right) \le C' t^{-(\kappa l)/2}. \tag{3.14}$$

Since, by construction, $l > 2/\kappa$, (3.13) and (3.14) guarantee that for $n > N_0$,

$$E\left(\sum_{i=0}^{lq-1} Y_{n,i}^2\right)^{-q} \le s + C' \int_s^\infty t^{-(\kappa l)/2} dt < \infty,$$

which yields (3.9). □

Lemma 4 below shows that (3.4) is easily found in many time series applications.

Lemma 4. *If ω_t's are i.i.d. random variables satisfying $E(\omega_1) = 0, E(\omega_1^2) = \sigma_\omega^2 > 0$, and $E(|\omega_1|^\alpha) < \infty$ for some $\alpha > 2$. Assume also that for some positive constant $M_0 < \infty$,*

$$\int_{-\infty}^{\infty} |\varphi(t)| dt \le M_0, \tag{3.15}$$

where $\varphi(t) = E(e^{it\omega_1})$ *is the characteristic function of* ω_1*. Then, for all* $-\infty < t < \infty$*,* $m \geq 1$ *and* $\|\mathbf{a}_m\| = 1$*, there is a finite positive constant* M_1 *such that*

$$\sup_{-\infty<x<\infty} f_{t,m,\mathbf{a}_m}(x) < M_1, \tag{3.16}$$

where $f_{t,m,\mathbf{a}_m}(\cdot)$ *is the density function of* $\sum_{j=1}^m a_j\omega_{t+1-j}$*. As a result, (3.4) follows.*

Proof. The proof is inspired by the ideas of Feller ([7], p. 516), which deal with the special case, $a_j = m^{-1/2}$ for all $j = 1, \ldots, m$. Without loss of generality, assume $\sigma_\omega^2 = 1$. Denote $Y = \sum_{j=1}^m a_j\omega_{t+1-j}$. Then, $\varphi_Y(t) = E(e^{itY}) = \prod_{j=1}^m \varphi_j(a_jt)$. By Chow and Teicher ([6], Theorem 8.4.1),

$$\varphi(a_jt) = 1 - \frac{a_j^2t^2}{2} + o(a_j^2t^2),$$

as $a_j^2t^2 \to 0$. This gives for $|a_jt| < \delta_1^*$, where δ_1^* is some small positive constant,

$$|\varphi(a_jt)| \leq 1 - \frac{a_j^2t^2}{4}. \tag{3.17}$$

On the other hand, since (3.15) yields $|\varphi(t)| \to 0$ as $|t| \to \infty$, by Chow and Teicher ([6], Corollary 8.4.2), $|\varphi(t)| < 1$ for all $t \neq 0$, and hence for all $|t| \geq \delta_1^*$ (with δ_1^* defined above),

$$|\varphi(t)| < \theta_1, \tag{3.18}$$

where θ_1 is some positive constant < 1. Now, by (3.17),

$$\begin{aligned}\int_{-\infty}^{\infty}\prod_{j=1}^m|\varphi(a_jt)|dt &\leq \int_{|t|<\frac{\delta_1^*}{O_m}} e^{\frac{-t^2}{4}}dt + \int_{|t|\geq\frac{\delta_1^*}{O_m}}\prod_{j=1}^m|\varphi(a_jt)|dt \\ &\leq \int_{-\infty}^{\infty} e^{\frac{-t^2}{4}}dt + \int_{|t|\geq\frac{\delta_1^*}{O_m}}\prod_{j=1}^m|\varphi(a_jt)|dt,\end{aligned} \tag{3.19}$$

where O_j is a permutation of $|a_j|$ satisfying $O_m \geq O_{m-1} \geq \cdots \geq O_1$. For $t \geq \delta_1^*/O_m$, (3.17), (3.18) and the fact that

$$\theta_1 = 1 - (1-\theta_1) \leq 1 - \frac{4(1-\theta_1)}{\delta_1^{*^2}}\frac{a_j^2\delta_1^{*^2}}{4O_m^2}$$

imply

$$|\varphi(a_jt)| \leq \max\{1 - \frac{a_j^2t^2}{4}, \theta_1\} \leq \max\{1 - \frac{a_j^2\delta_1^{*^2}}{4O_m^2}, \theta_1\} \leq 1 - \xi\frac{a_j^2\delta_1^{*^2}}{4O_m^2}, \tag{3.20}$$

where $0 < \xi < \min\{1, 4(1-\theta_1)/\delta_1^{*^2}\}$. In view of (3.20) and the fact that $\sum_{j=1}^{m-1} O_j^2 = 1 - O_m^2$,

$$\begin{aligned}\int_{|t|\geq\frac{\delta_1^*}{O_m}}\prod_{j=1}^m|\varphi(a_jt)|dt &\leq \frac{1}{O_m}\int_{-\infty}^{\infty} e^{-\frac{\xi\delta_1^{*^2}}{4O_m^2}\sum_{j=1}^{m-1}O_j^2}|\varphi(t)|dt \\ &= e^{\frac{\xi\delta_1^{*^2}}{4}}\frac{1}{O_m}e^{\frac{-\xi\delta_1^{*^2}}{4O_m^2}}\int_{-\infty}^{\infty}|\varphi(t)|dt \\ &\leq e^{\frac{\xi\delta_1^{*^2}}{4}}\sup_{x\geq1} xe^{\frac{-\xi\delta_1^{*^2}x^2}{4}}M_0 < \infty.\end{aligned} \tag{3.21}$$

By (3.21), (3.19), and the fact that

$$\sup_{-\infty<x<\infty} f_{t,m,\mathbf{a}_m}(x) \le \frac{1}{2\pi}\int_{-\infty}^{\infty}\prod_{j=1}^{m}|\varphi(a_j t)|dt,$$

(3.16) follows. In addition, it is not difficult to see that (3.4) can be deduced from (3.16). □

In the following lemma, some moment bounds for $(1/\sqrt{n})x_n$ and $(1/n)\sum_{i=1}^{n-1} x_i \times \varepsilon_{i+1}$, are obtained.

Lemma 5. *Assume models (1.1) and (1.2), with* $\sup_t E(|\varepsilon_t|^q) < \infty$ *and* $\sup_t E(|\omega_t|^q) < \infty$, *for some* $q \ge 2$. *Then,*

$$\text{(i)} \qquad \sup_{n\ge 1} E\left(\left|\frac{1}{\sqrt{n}}x_n\right|^q\right) < \infty, \tag{3.22}$$

$$\text{(ii)} \qquad \sup_{n\ge 1} E\left(\left|\frac{1}{n}\sum_{i=1}^{n-1} x_i\varepsilon_{i+1}\right|^q\right) < \infty. \tag{3.23}$$

Proof. The proof of Lemma 5 is similar to that of Ing ([9], Lemma 1). The details are omitted. □

Armed with the previous results, (3.2) is proved in the following theorem.

Theorem 2. *Assume that (1.1), (1.2), (3.4),* $\sup_t E(|\varepsilon_t|^q) < \infty$, *and* $\sup_t E(|\omega_t|^q) < \infty$ *are satisfied, where* $q > 4$. *Then, (3.2) holds. If we further assume that* $E(\varepsilon_t\omega_t) = \pi$ *is a constant independent of* t, *then (3.3) follows.*

Proof. By Lemmas 3 and 5, (3.1), and an argument similar to the one used in [9], Theorem 1, the claimed results can be obtained. □

The FPE of the least squares predictor is obtained in Corollary 1 below.

Corollary 1. *Assume that (2.7) and all assumptions of Theorem 2 hold. Then, (1.9) follows.*

Proof. By (2.15), (3.2), and Minkowski's inequality,

$$\lim_{n\to\infty}\frac{1}{\log n}\sum_{i=m^*}^{n} E\{x_{i-1}^2(\hat{\beta}_{i-1}-\beta)^2\} = 2\sigma^2, \tag{3.24}$$

where m^* is some positive integer independent of n. Now, (1.9) is guaranteed by (3.3) and (3.24). □

Corollary 1 and Theorem 1 together indicate an interesting result that the term of order $\log n$ in the APE and the term of order n^{-1} in the FPE share the same constant, $2\sigma^2$. For applications of this type of results to model selection problems, see [11]. Corollary 1 also shows that the FPE of the least squares predictor is not affected by the contemporary correlation between ε_t and ω_t. This is a somewhat unexpected feature because the least squares estimate itself does not possess this property. More specifically, by direct calculations, we have

$$n(\hat{\beta}_n - \beta) \Longrightarrow \frac{1}{\lambda}\frac{\rho\sigma_\omega\int_0^1 w_a(t)dw_a(t) + \sigma_\theta\int_0^1 w_a(t)dw_b(t)}{\int_0^1 w_a^2(t)dt}, \tag{3.25}$$

and

$$n^2(\hat{\beta}_n-\beta)^2 \Longrightarrow \frac{1}{\lambda^2}\frac{\left(\rho\sigma_\omega\int_0^1 w_a(t)dw_a(t)+\sigma_\theta\int_0^1 w_a(t)dw_b(t)\right)^2}{\left(\int_0^1 w_a^2(t)dt\right)^2}, \tag{3.26}$$

where λ is defined in Remark 2. By (3.26), an argument similar to that used in the proof of Theorem 2, and some algebraic manipulations,

$$\begin{aligned}\lim_{n\to\infty} n^2E(\hat{\beta}_n-\beta)^2 &= E\left\{\frac{1}{\lambda^2}\frac{\left(\rho\sigma_\omega\int_0^1 w_a(t)dw_a(t)+\sigma_\theta\int_0^1 w_a(t)dw_b(t)\right)^2}{\left(\int_0^1 w_a^2(t)dt\right)^2}\right\}\\ &= \frac{\rho^2}{\iota^2}E\left(\frac{\int_0^1 w_a(t)dw_a(t)}{\int_0^1 w_a^2(t)dt}\right)^2+\frac{\sigma_\theta^2}{\iota^2\sigma_\omega^2}E\left(\frac{1}{\int_0^1 w_a^2(t)dt}\right),\end{aligned} \tag{3.27}$$

where $\iota^2=\lambda^2\sigma_\omega^{-2}$. Ing ([9], (4.3)) showed that

$$E\left(\frac{\int_0^1 w_a(t)dw_a(t)}{\int_0^1 w_a^2(t)dt}\right)^2 \doteq 13.3. \tag{3.28}$$

By (3.6.4) and (3.6.5) of Arató and using a numerical integration method,

$$E\left(\frac{1}{\int_0^1 w_a^2(t)dt}\right) \doteq 5.6. \tag{3.29}$$

Consequently, (3.27)-(3.29) imply

$$\lim_{n\to\infty} n^2E(\hat{\beta}_n-\beta)^2 \doteq \frac{\rho^2}{\iota^2}13.3+\frac{\sigma_\theta^2}{\iota^2\sigma_\omega^2}5.6, \tag{3.30}$$

which obviously varies with the strength of dependence between ε_t and ω_t. In particular, if $\sigma^2=\sigma_\omega^2$, then $\rho=\mathrm{corr}(\varepsilon_t,\omega_t)$ and (3.30) can be rewritten as

$$\lim_{n\to\infty} n^2E(\hat{\beta}_n-\beta)^2 \doteq \frac{1}{\iota^2}[\rho^2 13.3+(1-\rho^2)5.6]. \tag{3.31}$$

As observed in (3.31), the larger the magnitude of the correlation between ε_t and ω_t is, the larger the mean squared error of the least squares estimate is, a result new to the literature.

As a final remark, we note that the square of the normalized estimate, $n^2(\hat{\beta}_n-\beta)^2$, and the square of normalized regressor, x_n^2/n, are not asymptotically uncorrelated. To see this, observe that $\lim_{n\to\infty}E(x_n^2/n)=\lambda^2$, which together with (3.30) and Corollary 1, gives

$$\begin{aligned}\lim_{n\to\infty} E\left(\frac{x_n^2}{n}\right)E\left\{n^2(\hat{\beta}_n-\beta)^2\right\} &\doteq 13.3\rho^2\sigma_\omega^2+5.6\sigma_\theta^2\\ &= 5.6\sigma^2+7.7\rho^2\sigma_\omega^2>2\sigma^2\\ &= \lim_{n\to\infty}E\left\{\frac{x_n^2}{n}n^2(\hat{\beta}_n-\beta)^2\right\}.\end{aligned}$$

Therefore, x_n^2/n and $n^2(\hat{\beta}_n-\beta)^2$ are (asymptotically) negatively correlated, which suggests that larger variation of x_n can yield a better estimation result. It is worth

mentioning that this special feature does not exist for the (asymptotically) stationary regressor. For example, when $x_t = \varsigma x_{t-1} + \eta_t$, with $|\varsigma| < 1$, following an argument used in Ing [10], it can be shown that

$$\lim_{n\to\infty} E(x_n^2) E\left\{[\sqrt{n}(\hat{\beta}_n - \beta)]^2\right\} = \lim_{n\to\infty} E\left\{x_n^2 n(\hat{\beta}_n - \beta)^2\right\} = \sigma^2.$$

Therefore, the square of the normalized estimate, $n(\hat{\beta}_n - \beta)^2$, and the square of the (normalized) regressor, x_n^2, are asymptotically uncorrelated in this case.

References

[1] Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243–247.

[2] Arató, M. (1982). *Linear Stochastic Systems With Constant Coefficients: A Statistical Approach.* Springer-Verlag, New York.

[3] Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Method.* Springer-Verlag, New York.

[4] Chan, N. H. and Wei, C. Z. (1988). Limiting distribution of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16** 367–401.

[5] Chow, Y. S. (1965). Local convergence of martingale and the law of large numbers. *Ann. Math. Statist.* **36** 552–558.

[6] Chow, Y. S. and Teicher, H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed. Springer, New York.

[7] Feller, W. (1971). *An Introduction to Probability Theory and Its Application*, Vol. II. Wiley, New York.

[8] Hamilton, J. D. (1994). *Time Series Analysis.* Princeton University Press, Princeton, NJ.

[9] Ing, C. K. (2001). A note on mean-squared prediction errors of the least squares predictors in random walk models. *J. Time Ser. Anal.* **22** 711–724.

[10] Ing, C. K. (2003). Multistep prediction in autoregressive processes. *Economet. Theory* **19** 254–279.

[11] Ing, C. K. (2004). Selecting optimal multistep predictors for autoregressive processes of unknown order. *Ann. Statist.* **32** 693–722.

[12] Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with application to identification and control systems. *Ann. Statist.* **10** 154–166.

[13] Móricz, F. (1976). Moment inequalities and the strong laws of large numbers. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **35** 299–314.

[14] Rissanen, J. (1986). Order estimation by accumulated prediction errors. In *Essays in Time Series and Applied Processes*, Special vol. **23A** of *J. Appl. Prob.* 55–61.

[15] Wei, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with application to time series. *Ann. Statist.* **15** 1667–1682.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 72–92

DOI: 10.1214/074921706000000969

Forecasting unstable processes

Jin-Lung Lin[1] **and Ching-Zong Wei**[2]

Academia Sinica

Abstract: Previous analysis on forecasting theory either assume knowing the true parameters or assume the stationarity of the series. Not much are known on the forecasting theory for nonstationary process with estimated parameters. This paper investigates the recursive least square forecast for stationary and nonstationary processes with unit roots. We first prove that the accumulated forecast mean square error can be decomposed into two components, one of which arises from estimation uncertainty and the other from the disturbance term. The former, of the order of $\log(T)$, is of second order importance to the latter term, of the order T. However, since the latter is common for all predictors, it is the former that determines the property of each predictor. Our theorem implies that the improvement of forecasting precision is of the order of $\log(T)$ when existence of unit root is properly detected and taken into account. Also, our theorem leads to a new proof of strong consistency of predictive least squares in model selection and a new test of unit root where no regression is needed.

The simulation results confirm our theoretical findings. In addition, we find that while mis-specification of AR order and under-specification of the number of unit root have marginal impact on forecasting precision, over-specification of the number of unit root strongly deteriorates the quality of long term forecast. As for the empirical study using Taiwanese data, the results are mixed. Adaptive forecast and imposing unit root improve forecast precision for some cases but deteriorate forecasting precision for other cases.

1. Introduction

Forecasting future observations is one of the major purpose of building a time series model. Even for the purpose of time series controlling, forecasting provide the essential basis. For this purpose, autoregressive (AR) models are widely used for their simplicity. For an AR(p) process,

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \epsilon_t \tag{1}$$

where $\phi(z) = 1 - \beta_1 z - \cdots - \beta_p z^p$ the characteristic polynomial determines the properties of the series. y_t is called stationary or stable if all roots of ϕ are outside the unit circle, unstable or nonstationary if some roots of ϕ are on the unit circle and explosive if some roots of ϕ are inside the unit circle. Previous analysis on forecasting theory either assume knowing true β_s' or only consider the stationary cases. For examples, Ing [8, 9] and Bhansali [1, 2] analyze the multistep prediction of stationary AR processes while Ing [7] derives the mean squares prediction errors of the least squares predictors in random walk model. Not much are known on

[1]Institute of Economics, Academia Sinica, 128 Sec. 2, Academia Rd., Nankang, Taipei, Taiwan 11529, e-mail: jlin@econ.sinica.edu.tw

[2]Institute of Statistical Science, Academia Sinica, 128 Sec. 2, Academia Rd., Nankang, Taipei, Taiwan 11529.

AMS 2000 subject classifications: primary 62G25; secondary 62M20.

Keywords and phrases: unit root, unstable process, adaptive forecast, direct forecast, plug-in forecast, strong convergence.

the forecasting theory for unstable process with estimated parameters. This paper investigates the recursive least square forecast for stable and unstable processes.

Let $\hat{y}_t$ be the forecast of y_t based upon information up to $t-1$. If one is interested in one-period forecast, $(y_t-\hat{y}_t)^2$ is the cost to be minimized. However, there are two situations where the accumulated cost function, $\sum_{k=1}^{t}(y_k-\hat{y}_k)^2$ is more appropriate. First, in the sequential forecast case, (see Goodwin and Sin [6]) the forecaster are updated sequentially over many periods and the accumulated cost function is the target to be minimized. Second, for a single realization of time series, the averaged accumulated cost function is often used as the yardstick to evaluate the out-of-sample forecasting performance of alternative forecasters.

Ing [7] advocated adopting the accumulated cost function $\sum_{t=1}^{T} E(y_t-\hat{y}_t)^2$ over the one-period expected loss function $E(y_{T+1}-\hat{y}_{T+1})^2$. For an AR(1) process, these two quantities are respectively:

$$\frac{1}{T-2}\sum_{t=3}^{T} E(y_t-\hat{y}_t)^2 = \sigma^2 + \frac{2\sigma^2\log(T)}{T} + o\left(\frac{\log(T)}{T}\right)$$

$$E(y_{T+1}-\hat{y}_{T+1})^2 = \sigma^2 + \frac{2\sigma^2}{T} + o\left(\frac{1}{T}\right)$$

when true $\beta_1 = 1$. In other words, the efficiency loss for not taking the unit root into consideration is greater for the accumulated cost function than the one-period cost function. See also Ing and Wei [11]. It is worth mentioning that Rissanen [14] predictive least square (PLS) for model selection built upon accumulated cost function minimization. See also Wei [18].

Under the assumption that $E(\epsilon_t^2|\mathcal{F}_{t-1}) = \sigma^2$ a.s. for all t, where $\mathcal{F}_{t-1}$ is the sigma field generated by $\{x_s, s \le t-1\}$, then it can be shown that under appropriate assumptions that $\frac{1}{T}\sum_{t=1}^{T}(y_t-\hat{y}_t)^2 \longrightarrow \sigma^2 \quad a.s.$ But by Chow [4], it is seen that

$$\sum_{t=1}^{T}(y_t-\hat{y}_t)^2 = \sum_{t=1}^{T}\epsilon_t^2 + C_T(1+o(1)) \quad a.s. \quad \text{on the set } \{C_T \to \infty\}$$

$$\sum_{t=1}^{T}(y_t-\hat{y}_t)^2 = \sum_{t=1}^{T}\epsilon_t^2 + C_T(1+O(1)) \quad a.s. \quad \text{on the set } \{C_T < \infty\}$$

where

$$C_T = \sum_{t=1}^{T}(y_t-\hat{y}_t-\epsilon_t)^2$$

While $\sum_{t=1}^{T}\epsilon_t^2$ is larger in order than C_T, it is common for all forecasters and cannot be removed. Hence C_T becomes a more important quantity when evaluating the performance of alternative forecasters.

Let $\hat{\beta}_t$ be the least square estimate of β

$$\hat{\beta}_t = [\sum_{k=1}^{t}\boldsymbol{Y}_{k-1}\boldsymbol{Y}'_{k-1}]^{-1}\sum_{k=1}^{t}\boldsymbol{Y}'_{k-1}y_k$$

where $Y_t = \{y_1,\ldots,y_t\}'$, then $\hat{y}_t = \hat{\beta}'_{t-1}Y_{t-1}$ is the least square prediction of y_t at time $t-1$.
Let

$$\phi(z) = (z-1)^a(z+1)^b\Pi_{k=1}^{l}(z^2-2\cos\theta_k z+1)^{d_k}\pi(z)$$

where all roots of $\pi(z)$ are all outside the unit circle. Wei [17] proves that,

$$C_T \to (p + a^2 + b^2 + 2\sum_{k=1}^{l} d_k^2)\sigma^2 \log(T) \quad \text{in probability.} \tag{2}$$

In other words, when $\phi(z)$ has multiple unit roots the accumulated loss increase not linearly with the number of unit roots but at the rate of the square of the number of unit roots.

In this paper, we prove that when $\phi(z)$ has no complex roots, the convergence in (2) can be improved to be almost surely. This result could lead to a new proof of strong consistency of PLS in AR model selection. It is also conjectured that the result of almost surely convergence hold for the case of complex unit roots. We conduct several simulation experiments to assess the convergence result for various sample sizes. In addition, we also consider the impact of near unit root and model mis-specification on multi-step forecasting. Finally, we apply our methods to six real macroeconomic series in Taiwan. Forecasting performance of various forecasters and adaptive forecaster are investigated.

The rest of the paper is organized as follows. The proof of the main theorem is put in Section 2. Section 3 illustrates implications and applications of our main theorem. Section 4 discusses multi-step and adaptive forecast. Monte Carlo results are reported in Section 5 and Section 6 summarizes the empirical results. Section 7 concludes.

2. Main theorem

Assume that ϵ_t are *i.i.d.* random variables with $E(\epsilon_t) = 0$ and $0 < E(\epsilon_t^2) = \sigma^2 < \infty$. Let $\boldsymbol{X}_t = (x_{t-1}, \ldots, x_{t-p})'$, $S_T = \sum_{t=1}^{T} \epsilon_t$ and $T_T = (-1)^T \sum_{t=1}^{T} (-1)^t \epsilon_t = \epsilon_t + (-1)T_{T-1}$.

Lemma 1. *Assume that* $\boldsymbol{X}_{t+1} = A\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t$ *, where* $\boldsymbol{\epsilon}_t = (\epsilon_t, 0, \ldots, 0)'$ *and the eigenvalues of* A *are all inside the unit circle. Then*

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} \boldsymbol{X}_t S_t}{\sqrt{T\sum_{t=1}^{T} S_t^2}} = 0 \quad a.s.$$

and

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} \boldsymbol{X}_t T_t}{\sqrt{T\sum_{t=1}^{T} T_t^2}} = 0 \quad a.s.$$

Proof. It is known from Lai and Wei [12] [pages 363 and 364] that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{X}_t \boldsymbol{X}_t' = \Sigma \quad a.s. \tag{3}$$

where Σ is a positive definite matrix,

$$\limsup_{T\to\infty} \frac{\sum_{t=1}^{T} S_t^2}{T^2 \log\log(T)} = \frac{8\sigma^2}{\pi^2} \quad a.s. \tag{4}$$

and

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T} S_t^2}{T^2/\log\log(T)} = \frac{\sigma^2}{4} \quad a.s. \tag{5}$$

Let $\|u\|$ denote the Euclidean norm of a k-dimensional vector $u = (u_1, \ldots, u_k)'$, i.e., $\|u\|^2 = \sum_{i=1}^{k} u_i^2$. By (3), $\frac{\|\boldsymbol{X}_T\|^2}{T} \to 0 \quad a.s.$ and in turn we have that

$$0 \le \boldsymbol{X}_T'(\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t')^{-1}\boldsymbol{X}_T \le \frac{\|\boldsymbol{X}_T\|^2}{\lambda_{\min}(\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t')} = \frac{\|\boldsymbol{X}_T\|^2/T}{\lambda_{\min}(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t')} \to 0 \ a.s.$$

and

$$\boldsymbol{X}_T'(\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t')^{-1}\boldsymbol{X}_T \to 0 \quad a.s. \tag{6}$$

where $\lambda_{\min}(A)$ denotes the minimal eigenvalue of matrix A.

Furthermore, by the law of iterative logarithm,

$$\limsup_{T\to\infty} \frac{S_T^2}{2T\log\log T} = \sigma^2 \quad a.s.$$

Hence (5) implies that

$$\begin{aligned} \frac{S_T^2}{\sum_{t=1}^{T} S_t^2} &= O\left(\frac{T\log\log T}{T^2/\log\log T}\right) \\ &= O\left(\frac{(\log\log T)^2}{T}\right) \\ &= o(1) \quad a.s. \end{aligned} \tag{7}$$

Now, let

$$\boldsymbol{Z}_T = \frac{\sum_{t=1}^{T} \boldsymbol{X}_t S_t}{\left(T\sum_{t=1}^{T} S_t^2\right)^{1/2}}.$$

Then

$$\begin{aligned} \boldsymbol{Z}_T - \boldsymbol{Z}_{T-1} &= \boldsymbol{Z}_T - \frac{\sum_{t=1}^{T-1} \boldsymbol{X}_t S_t}{(T\sum_{t=1}^{T} S_t^2)^{1/2}} - \boldsymbol{Z}_{T-1}\left(1 - (\frac{(T-1)\sum_{t=1}^{T-1} S_t^2}{T\sum_{t=1}^{T} S_t^2})^{1/2}\right) \\ &= \frac{\boldsymbol{X}_T S_T}{(T\sum_{t=1}^{T} S_t^2)^{1/2}} - \boldsymbol{Z}_{T-1}\left(1 - (\frac{T-1}{T} - \frac{T-1}{T}\frac{S_T^2}{\sum_{t=1}^{T} S_t^2})^{1/2}\right) \\ &= \frac{\boldsymbol{X}_T S_T}{(T\sum_{t=1}^{T} S_t^2)^{1/2}} - \boldsymbol{Z}_{T-1}\, o(1), \quad \text{by (7)} \\ &= o(1) - o(1), \quad \text{since} \quad \sup_T \|\boldsymbol{Z}_T\| \le \{\frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{X}_t\|^2\}^{1/2} \quad a.s. \\ &= o(1) \end{aligned} \tag{8}$$

But,

$$\begin{aligned}\sum_{t=1}^{T} \boldsymbol{X}_t S_t &= \sum_{t=1}^{T} (A\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t) S_t \\ &= A\sum_{t=1}^{T} \boldsymbol{X}_{t-1} S_{t-1} + A\sum_{t=1}^{T} \boldsymbol{X}_{t-1}\epsilon_t + \sum_{t=1}^{T} \boldsymbol{\varepsilon}_t S_{t-1} + \sum_{t=1}^{T} \epsilon_t^2 \\ &= A\sum_{t=1}^{T} \boldsymbol{X}_{t-1} S_{t-1} + o((\sum_{t=1}^{T} \|\boldsymbol{X}_{t-1}\|^2)^{1/2} (\log \sum_{t=1}^{T} \|\boldsymbol{X}_{t-1}\|^2)^{\frac{1+\sigma}{2}}) \\ &\quad + o((\sum_{t=1}^{T} S_{t-1}^2)^{1/2} (\log \sum_{t=1}^{T} S_{t-1}^2)^{\frac{1+\sigma}{2}}) + O(T) \quad a.s. \\ &= A(\sum_{t=1}^{T} \boldsymbol{X}_{t-1} S_{t-1}) + o(T^{1/2} (\log T)^{\frac{1+\sigma}{2}}) \\ &\quad + o((\sum_{t=1}^{T} S_{t-1}^2)^{1/2} (\log T)^{\frac{1+\sigma}{2}}) + O(T)\end{aligned}$$

This implies that

$$\boldsymbol{Z}_T = A\boldsymbol{Z}_{T-1}(1 + o(1)) + o(1) \quad a.s. \tag{9}$$

Combining (8) and (9), we have that

$$\boldsymbol{Z}_{T-1} - A\boldsymbol{Z}_{T-1} = o(1) \quad a.s. \tag{10}$$

Therefore, any limit point z of $\{z_T\}$ would satisfy

$$\boldsymbol{Z} - A\boldsymbol{Z} = 0 \tag{11}$$

Since 1 is not an eigenvalue of A, $\boldsymbol{Z} = 0$. Using the same method one can prove that

$$\frac{\sum_{t=1}^{T} \boldsymbol{X}_t T_t}{(T\sum_{t=1}^{T} T_t^2)^{1/2}} = 0 \quad a.s.$$

This proves Lemma 1. □

Lemma 2. *If $E|\epsilon_t^\alpha| < \infty$ for some $\alpha > 2$, then*

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} S_t T_t}{\sqrt{\sum_{t=1}^{T} S_t^2 \sum_{t=1}^{T} T_t^2}} = 0 \quad a.s.$$

Proof. Note that $\tilde{T}_T = (-1)^T T_T = \sum_{t=1}^{T} (-1)^T \epsilon_t$. Using theorem 3.2 of Phillip in page 234 of Eberlein and Taqqu [5], (4) and (5) hold if we replace S_t by T_t.

Therefore,

$$\frac{T_T^2}{\sum_{t=1}^{T} T_t^2} = \frac{\tilde{T}_T^2}{\sum_{t=1}^{T} \tilde{T}_t^2} \to 0 \quad a.s.$$

Let

$$u_T = \frac{\sum_{t=1}^{T} S_t T_t}{\sqrt{\sum_{t=1}^{T} S_t^2 \sum_{t=1}^{T} T_t^2}}.$$

Then

$$\begin{aligned} u_T - u_{T-1} &= \frac{S_T T_T}{\sqrt{\sum_{t=1}^{T} S_t^2 \sum_{t=1}^{T} T_t^2}} + u_{T-1}\left(\sqrt{\frac{\sum_{t=1}^{T-1} S_t^2 \sum_{t=1}^{T} T_t^2}{\sum_{t=1}^{T} S_t^2 \sum_{t=1}^{T} T_t^2}} - 1\right) \\ &= o(1) \quad a.s. \end{aligned} \tag{12}$$

But

$$\begin{aligned} \sum_{t=1}^{T} S_t T_t &= \sum_{t=1}^{T} (S_{t-1} + \epsilon_t)(-T_{t-1} + \epsilon_t) \\ &= -\sum_{t=1}^{T} S_{t-1} T_{t-1} + \sum_{t=1}^{T} S_{t-1}\epsilon_t - \sum_{t=1}^{T} T_{t-1}\epsilon_t + \sum_{t=1}^{T} \epsilon_t^2 \\ &= -\sum_{t=1}^{T-1} S_t T_t + o((\sum_{t=1}^{T} S_{t-1}^2)^{1/2})(\log(\sum_{t=1}^{T} S_{t-1}^2)) \\ &\quad + o((\sum_{t=1}^{T} T_t^2)^{1/2}(\log(\sum_{t=1}^{T} T_t^2))) + O(T) \quad a.s. \end{aligned}$$

Therefore,

$$\begin{aligned} u_T &= -\frac{\sum_{t=1}^{T-1} S_t T_t}{\sqrt{\sum_{t=1}^{T} S_t^2}\sqrt{\sum_{t=1}^{T} T_t^2}} + o(\frac{\log(\sum_{t=1}^{T} S_{t-1}^2)}{\sqrt{\sum_{t=1}^{T} T_t^2}}) + o(\frac{\log(\sum_{t=1}^{T} T_t^2)}{\sqrt{\sum_{t=1}^{T} S_t^2}}) \\ &\quad + o(\frac{T}{\sqrt{\sum_{t=1}^{T} T_t^2 \sum_{t=1}^{T} S_t^2}}) \\ &= -U_{T-1}(1 + o(1)) + o(1) \quad a.s. \\ &= -u_{T-1} + o(1) \quad a.s. \end{aligned} \tag{13}$$

Combining (12) and (13), since

$$u_T = o(1) \quad a.s. \quad u_T \longrightarrow 0 \quad a.s. \qquad \square$$

Now, we are ready to state our main result.
Let

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \epsilon_t \tag{14}$$

be an AR(p) model with

$$\phi(z) = 1 - \beta_1 z - \cdots - \beta_p z^p \tag{15}$$

$$= (1 - z)(1 + z)\Psi(z) \tag{16}$$

where $\Psi(z) = 1 - \Psi_1 z - \cdots - \Psi_q z^q$ is a polynomial of order $q = p - 2$ which has all roots outside the unit circle.

Theorem 1. *Assume that the AR(p) model (14) satisfies (16). If $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with $E|\epsilon_t|^\alpha < \infty$, where $\alpha > 2$, and $y_0, \ldots, y_{1-p}$ is independent of $\{\epsilon_t\}$ then*

$$\lim_{T\to\infty} \frac{1}{\log T} \log\det\Big(\sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t'\Big) = (p+2) \quad a.s. \tag{17}$$

where $\boldsymbol{y}_t' = (y_t, \ldots, y_{t-p+1})$.

Proof. By Chan and Wei [3] there exists a non-singular $p \times p$ matrix Q such that $Q\boldsymbol{y}_t = (u_t, v_t, \boldsymbol{x}_t')$, where

$$\begin{aligned} \boldsymbol{x}_t &= (x_{t-1}, \ldots, x_{t-q})', \\ u_t &= u_{t-1} + \epsilon_t, \\ v_t &= -v_{t-1} + \epsilon_t \quad \text{and} \\ x_t &= \Psi_1 x_{t-1} + \cdots + \Psi_q x_{t-q}. \end{aligned}$$

Therefore, if we let $\boldsymbol{z}_t = Q\boldsymbol{y}_t$,

$$\det\Big(\sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t'\Big) = \det[Q^{-1} \sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t' Q^{-1}] = \frac{\det(\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t')}{(\det(Q))^2}.$$

To show (17), it is sufficient to show

$$\lim_{T\to\infty} \frac{1}{\log T} \log\det\Big(\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t'\Big) = (p+2) \quad a.s. \tag{18}$$

Let

$$G_T = \begin{pmatrix} (\sum_{t=1}^{T} u_t^2)^{-1/2} & 0 & 0 \\ 0 & (\sum_{t=1}^{T} v_t^2)^{-1/2} & 0 \\ 0 & 0 & T^{-1/2} I_q \end{pmatrix},$$

where I_q is the $q \times q$ identity matrix.

Then

$$G_T \sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t' G_T = \begin{pmatrix} 1 & a_T & \boldsymbol{b}_T' \\ a_T & 1 & \boldsymbol{c}_T' \\ \boldsymbol{b}_T & \boldsymbol{c}_T & \Gamma_T \end{pmatrix},$$

where

$$\begin{aligned} a_T &= \frac{(\sum_{t=1}^{T} u_t v_t)}{[(\sum_{t=1}^{T} u_t^2)(\sum_{t=1}^{T} v_t^2)]^{1/2}}, \\ \boldsymbol{b}_T &= \frac{\sum_{t=1}^{T} u_t \boldsymbol{x}_t}{(T\sum_{t=1}^{T} u_t^2)^{1/2}}, \\ \boldsymbol{c}_T &= \frac{\sum_{t=1}^{T} v_t \boldsymbol{x}_t}{(T\sum_{t=1}^{T} v_t^2)^{1/2}}, \end{aligned}$$

and

$$\Gamma_T = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'.$$

Let

$$A = \begin{pmatrix} \Psi_1 \dots \Psi_q \\ \mathbf{0} \quad I_{q-1} \end{pmatrix}.$$

Then A has all eigenvalues inside the unit circle and $\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t$. Therefore, there exist a non-singular matrix Γ such that

$$\lim_{T\to\infty} \Gamma_T = \Gamma \quad a.s.$$

Furthermore, by Lemma 1 and 2,

$$\lim_{T\to\infty} a_T = 0,$$
$$\lim_{T\to\infty} \boldsymbol{c}_T = \mathbf{0} \quad a.s.$$

Consequently,

$$\lim_{T\to\infty} G_T \sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t' G_T = \begin{pmatrix} 1\ 0\ \mathbf{0}' \\ 0\ 1\ \mathbf{0}' \\ \mathbf{0}\ \mathbf{0}\ \Gamma \end{pmatrix}$$

Since Γ is nonsingular, (18) is proved if

$$\begin{aligned} \log\det(G_T^{-2}) &= \log(\sum_{t=1}^{T} u_t^2) + \log(\sum_{t=1}^{T} v_t^2) + q\log T \\ &\sim (p+2)\log T \quad a.s. \end{aligned} \tag{19}$$

By (4) and (5) of Lemma 1,

$$\lim_{T\to\infty} \frac{1}{\log T} \sum_{t=1}^{T} u_t^2 = 2 \quad a.s.$$

Similar result holds for $\{v_t\}$. Therefore,

$$\log\det(G_T^{-2}) \sim (4+q)\log T = (p+2)\log T \quad a.s.$$

This completes our proof. □

Remark 1. Theorem 3 of Wei [17] shows that under similar assumptions as in our analysis,

$$C_T \sim \sigma^2 \log\det(\sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t') \quad a.s. \tag{20}$$

Thus,

$$C_T \sim (p+2)\sigma^2 \log(T) \quad a.s.$$

Remark 2. Theorem 1 and Remark 1 have an immediate implication for model selection and can greatly simplify the proof of Theorem 3.5 of Wei [18]. Let p^* be known and $p_0 = \max\{j : \beta_j \neq 0, 1 \leq j \leq p^*\}$ as in (1). Denote $PLS_T(p) = \sum_{t=t_0}^{T}(y_t - \hat{y}_t)^2$ where $\hat{y}_t$ is the forecast of y_t based upon information up to t-1 using the AR(p) model as in (1) and $PLS_T(\hat{p}_T) = \inf\{PLS_T(j) : 0 \leq j \leq p^*\}$. Wei [18] showed that for both cases of underspecifying and overspecifying AR order (j),

$\mathrm{P}(PLS_T(j) > PLS_T(p_0) \quad eventually) = 1$. Thus, $\mathrm{P}[\hat{p}_T = p_0 \quad eventually] = 1$. For the case of overspecification, Wei decomposed $\phi_p(z)$ into a sum of a unit root component and a stable component, and worked out the differnece of C_T between the true and the overspecified models. Our results can greatly simplify the proof. Let $C_T^{(j)} = \sum_{t=1}^{T}(y_t - \hat{y}_t^{(j)} - \epsilon_t)^2$ where $\hat{y}_t^{(j)}$ is the forecast of y_t at $t-1$ using the AR(j) model. For the case of overspecification, $\beta_j = 0, \forall j > p_0$. Applying Theorem 1 and Remark 1, $C_T^{(j)} \to (j+2)\sigma^2 \log(T) > (p_0+2)\sigma^2 \log(T) = C_T^{(p_0)} \quad a.s.$ As for the case of underspecification, $l < p_0$, the desired result, $\mathrm{P}[PLS_T(l) > PLS_T(p_0) \quad eventually] = 1$, is a direct consequence of Theorem 3.2 of Wei [18] since $\beta_{p_0} \neq 0$. Thus, $\mathrm{P}[\hat{p}_T = p_0 \quad eventually] = 1$.

3. Implications and applications of the main theorem

We have just proved that for an AR(p) process, $C_T = p\sigma^2 \log(T)$ if it is stationary and $C_T = (p+1)\sigma^2 \log(T)$ if there is an root of 1. Our theorem implies that if the existence of unit root is properly detected and unit root constraint is imposed in forming the forecast, then $C_T = (p-1)\sigma^2 \log(T)$. That is, for model with unit root, estimation is done for the differenced series rather than level of the series. By so doing, we reduce C_T by $2\sigma^2 \log(T)$ which could be substantial for large T and σ^2. However, it should be noted that $\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$ is not severely affected by existence of unit root since C_T, which is of the order of $\log(T)$, is dominated by $\sum_{t=1}^{T}\epsilon_t^2$, which is of the order T. This result is natural since it is the long term forecast and not the short term forecast that unit root has strong impact. These findings are further confirmed in our simulation study in Section 5.

In addition, our theorem implies that for AR(p) processes with root equal to or less than 1 in magnitude, as $T \longrightarrow \infty$,

$$\log\det \frac{1}{\log(T)}\Big(\sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t'\Big) \longrightarrow c \quad a.s. \tag{21}$$

where $c = (p+1)$ if there is a root of 1 and $c = p$ if all roots are less than one. Equivalently,

$$\hat{d}_T = \Big[\frac{1}{T}\log\det \sum_{t=1}^{T} \boldsymbol{y}_t \boldsymbol{y}_t' - p\Big]^{1/2} \longrightarrow d, \quad a.s. \tag{22}$$

where d is 1 if there is a root of 1 and 0 if there is no unit root. Note that if p is unknown but $r \geq p$ is given, (22) is still true with r replacing p in (22) and in the definition of $\boldsymbol{y}_t$ in (17). In other words, our theorem proves that $\hat{d}_T$ can be used as a test statistic for unit root. This issue will be further investigated in future research.

4. Multi-step and adaptive forecast

Our previous analysis focuses on 1-step forecast and there are cases when multiple-step forecast is the main concern. It is conjectured that our results can be extended to multi-step forecast but the issue will be pursued elsewhere. Instead, we shall concentrate our discussion on the relationship between model misspecification and adaptive forecast.

By (1), we have

$$y_{t+h} = \beta_1 y_{t+h-1} + \cdots + \beta_p y_{t+h-p} + \epsilon_{t+h} \tag{23}$$

and

$$\hat{y}_{t+h} = \hat{\beta}_1 \hat{y}_{t+h-1} + \cdots + \hat{\beta}_p \hat{y}_{t+h-p} \tag{24}$$

where $\hat{y}_{t+h-k} = y_t$ for $h \leq k$. So, (24) can be recursively solved in the order of $\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots, \hat{y}_{t+h}$. This is the conventional Box-Jenkins multi-step forecaster.

Another way of generating the multi-step forecast is to solve the model that minimizes the multi-step forecast error and then use it to form multi-step forecast (see Ing [8], Bhansali [2], Weiss [19], and Tiao and Tsay [15]). More specifically, the h-step forecast error $e_t(h)$ at time t is

$$e_t(h) = \epsilon_{t+h} + \Psi_1 \epsilon_{t+h-1} + \cdots + \Psi_{h-1} \epsilon_{t+1}$$

where Ψ_i is defined by $[1 - \beta B - \cdots - \beta_p B^p]^{-1} = \Psi_0 + \Psi_1 B + \cdots$. The cost function to be minimized is

$$C(h) = \sum_{t=1}^{T-h} e_t^2(h) \tag{25}$$

Note that for different h different models are used and this explains the name 'adaptive' forecast. Solving (25) involves nonlinear optimization as Ψ_i is a nonlinear function of $(\beta_1, \ldots, \beta_p)$. In practice, approximate linear model is used. That is, the following regression is performed

$$y_t = a_1 y_{t-h} + a_2 y_{t-h-1} + \cdots + a_p y_{t-h-p+1} + b_t$$

and

$$\hat{y}_{t+h} = a_1 y_t + a_2 y_{t-1} + \cdots + a_p y_{t-p+1}$$

The idea behind the adaptive forecast is that if the model is misspecified, that is, p is mistakenly chosen, then this mistake will be amplified radically for the long term forecast. Adaptive forecast could avoid this compounding impact. It is reasonable to expect good performance of Box-Jenkins forecaster for the correctly specified model and good performance of adaptive forecaster for misspecified model.

Ing, Lin and Yu [10] propose a predictor selection criterion to choose the best combination of prediction models (AR lags) and prediction methods (adaptive or plug-in). When there is only one unit root, the proposed method is proved to be asymptotically efficient in the sense that the predictor converges with probability one to the optimal predictor which has minimal loss function.

5. Monte Carlo experiments

To assess the theoretical results obtained in previous section and acquire experience about empirical analysis in the sequel, we conduct two Monte Carol experiments. The first is to investigate the finite sample properties of C_T in theorem 1 and the second on forecast comparison between alternative forecasters. For both cases, we generate data from the following four models:

- Model 1: $(1-0.5B)^2(1-B)y_t = \epsilon_t$ or $y_t = 2y_{t-1}-1.25y_{t-2}+0.25y_{t-3}+\epsilon_t$. Roots are 0.5, 0.5 and 1.0 respectively.
- Model 2: $(1-0.5B)^2(1-.99B)y_t = \epsilon_t$ or $y_t = 1.99y_{t-1} - 1.24y_{t-2} + 0.2475y_{t-3} + \epsilon_t$
  Roots are 0.5, 0.5 and 0.99 respectively.
- Model 3: $(1-0.5B)^2(1-.95B)y_t = \epsilon_t$ or $y_t = 1.95y_{t-1} - 1.2y_{t-2} + 0.2375y_{t-3} + \epsilon_t$
  Roots are 0.5, 0.5 and 0.95 respectively.
- Model 4: $(1-0.5B)^3y_t = \epsilon_t$ or $y_t = 1.5y_{t-1} - 0.75y_{t-2} + 0.125y_{t-3} + \epsilon_t$
  All roots are 0.5.

σ^2 is set to be 1 for all models.

5.1. Monte Carlo experiment on C_T

The number of replications are 1000 for each experiment. For each, realization, 10 sets of samples are drawn from each model with sample size, T, varying from 100, 200 to 1000. For each sample, starting from $t = t_0(=10)$, the model parameters are estimated and is then used to forecast $t+1$. Then we reestimate the model using sample from 1 to $t+1$ and forecast $t+2$. The process is repeated until when $T-1$ sample is used to estimate the model and then used to forecast y_T. The forecast mean square error is then summed from t_0+1 to T to obtain $\hat{C}_T$. Finally, we compute the averaged $\hat{C}_T$ obtained from 1000 replications. In other words,

$$\hat{C}_T = \frac{\sum_{i=1}^{1000}\sum_{t=t_0}^{T-1}(\hat{y}_{i,t+1} - y_{i,t+1})^2}{(1000)(T-t_0)} \tag{26}$$

In addition, for each model, we repeat the procedure above with the constraint that one of the root is equal to one. The results are summarized in Table 1. As one can easily see, over 40 millions regressions have to performed to obtain this table and usage of updating formula can significantly reduce the computation burden. In Table 1, the first column is sample size. Results for first model with 0 unit root ($d = 0$) and 1 unit root ($d = 1$) are put in second and third columns. Results for the other three models are put in columns 4 to 9. Our theory predicts that: (1)

Table 1
C_T for simulated data

	Roots are							
	0.5,0.5,1.0		0.5,0.5,0.99		0.5,0.5,0.95		0.5,0.5,0.5	
T	$d=0$	$d=1$	$d=0$	$d=1$	$d=0$	$d=1$	$d=0$	$d=1$
100	23.47	12.33	23.47	12.33	23.80	15.36	21.07	23.22
200	27.55	14.71	27.55	14.71	27.71	20.19	24.28	37.60
300	29.90	16.06	29.90	16.06	29.83	23.90	26.09	50.86
400	31.49	17.00	31.49	17.00	31.21	27.17	27.32	63.57
500	32.75	17.73	32.75	17.73	32.26	30.27	28.29	75.96
600	33.76	18.30	33.76	18.30	33.09	33.12	29.04	88.12
700	34.62	18.79	34.62	18.79	33.80	36.01	29.69	100.41
800	35.38	19.22	35.38	19.22	34.40	38.89	30.26	112.72
900	35.99	19.60	35.99	19.60	34.94	41.65	30.76	124.80
1000	36.55	19.94	36.55	19.94	35.42	44.37	31.21	136.93
β	5.2849	2.8583	5.2849	2.8583	5.1947	5.2064	4.5635	13.8536
R^2	0.9988	0.9902	0.9988	0.9902	0.9920	0.6315	0.9930	0.4471

TABLE 2
MSE for simulated Data

	Roots are							
	0.5,0.5,1.0		0.5,0.5,0.99		0.5,0.5,0.95		0.5,0.5,0.5	
T	$d=0$	$d=1$	$d=0$	$d=1$	$d=0$	$d=1$	$d=0$	$d=1$
100	117.81	106.92	117.81	106.92	118.33	110.06	115.59	118.17
200	227.10	214.61	227.10	214.61	227.36	220.24	224.11	237.66
300	334.88	321.53	334.88	321.53	334.97	329.57	331.53	356.59
400	441.30	427.33	441.30	427.33	441.27	437.73	437.66	474.10
500	547.43	533.00	547.43	533.00	547.19	545.69	543.55	591.33
600	653.42	638.61	653.42	638.61	653.01	653.67	649.36	708.93
700	759.51	744.24	759.51	744.24	758.92	761.63	755.18	826.46
800	865.20	849.45	865.20	849.45	864.35	869.13	860.55	943.18
900	970.92	954.95	970.92	954.95	970.01	976.96	966.16	1060.21
1000	1076.76	1060.56	1076.76	1060.56	1075.72	1084.89	1071.91	1177.63

$\hat{C}_T$ increases linearly with $\log(T-t_0)$ and (2) $\hat{C}_T$ without unit root constraint is 2 times $\hat{C}_T$ with unit root constraint.

We run a simple regression of $\hat{C}_T$ against $\log(T-t_0)$ without intercept for each model and report the regression coefficients and R^2 in the last row of Table 1. For column 2 and 3 of the table, the regression coefficients are 5.2849 and 2.8583 respectively while R^2 are greater than 0.99 for both cases. In summary, model 1 conforms the theoretical results.

As for model 2, one of the root is 0.99. Since it is the 1-step that is the main concern here, the result is almost the same with model 1. This is consistent with the findings of Lin and Tsay [13] that unit root or not does not matter much for short term forecast.

For model 3, the largest root is 0.95 which is not close to 1 enough. Imposing unit root constraint produces much larger $\hat{C}_T$ and the stable relationship between $\hat{C}_T$ and $\log(T)$ deteriorates greatly as is seen from poor R^2. This can be justified by the fact that differencing a stationary process produce a unit root in the MA component which can not be approximated by high order AR. The situation become much worse for model 4 where all roots are equal to 0.5.

For the purpose of comparison, we also report the corresponding conventional MSE ($\sum_{t=1}^{T}(y_t-\hat{y}_t)^2$) for the same 4 models above in Table 2. We observed from the table that contrary to the case for $\hat{C}_T$, the MSE for $d=0$ is about the same as for $d=1$. This confirms our previous analysis that C_T, though an important quantity for determining the quality of forecast, is of second order importance as compared to $\sum_{t=t_0+1}^{T}\epsilon_t^2$. For 1-step forecast the distinction between unit root and near unit root does not matter much.

5.2. Monte Carlo experiment on short-term and long-term forecast comparison

This simulation is designed to evaluate the short-term and long-term forecasting performance of alternative forecasters. The number of replications are again 1000. For each replication, 400 observations are generated from the four models above. The first 300 observations are reserved for estimation and then used to produce 1 to 60 steps forecast. Next, the model are re-estimated using the first 301 observations and then used to forecast 1 to 60 steps ahead. The procedure is repeated until when the first 399 observations is used for estimation and the last 1-step ahead

forecast is formed. So, we have 100 1-step forecasts, 99 2-step forecasts and 40 60-step forecasts. Then, we compute root mean square error (RMSE) for forecast of each step. Finally, the resulting RMSE is averaged over 1000 replications. More specifically, letting $\epsilon_{i,t}(k)$ be the k period ahead forecast error at time t of the i-th replication. Then

$$\text{RMSE}(\ell) = E(\ell) = \sqrt{\frac{\sum_{i=1}^{1000}\sum_{t=300}^{400-\ell}\epsilon_{i,t}^2(\ell)}{(1000)(100-\ell+1)}} \tag{27}$$

The simulation results are put in Tables 3 to 6. In each table, column 1 is steps of forecast, column 2 is the RMSE for model with $p = 3$ and $d = 0$, serving as the benchmark for forecast comparison. Columns 3 to 7 are $E(\ell)$ ratios of model with various p and d to column 2.

From these tables we observe the following. First, for stationary processes, the $E(\ell)$ for the correctly model converges to a constant with the rate of convergence depending upon the value of the root. For root of 0.5, the $E(\ell)$ approach a constant as early as $\ell = 6$ while for root of 0.95 ℓ does not stabilize until 30. As for root of .99, it is so close to 1 and $E(\ell)$ is still increasing after $\ell = 60$. For process with unit root $E(\ell)$ increases with ℓ for all the whole range of ℓ. Second, the true model outperforms other misspecified models in forecasting. Third, over-specification of unit results in poor forecast. For the case of model 4 (Table 6) $E(\ell)$ for $d = 1$ is 5% higher than $d = 0$ and jumps to more than 50% for ℓ greater than 40. For model 3, one of the root is 0.95 and the forecaster for $d = 1$ is still 45% worse than $d = 0$ though a little better than model 4. As for model 2, one of the root is 0.99 and for up to 20 steps, $d = 1$ fares as well as $d = 0$ and is only 10% worse than the true model at 60-step forecast. Fourth, under-specification of unit root only results in small increase of $E(\ell)$. From column 2 of Table 3, the inefficiency is less than 4% from 1-step to 60-step forecasts. Fifth, under- or over-specification of AR order

TABLE 3
Forecasting comparison for simulated data: true $p = 3$, roots are 0.5, 0.5, 1.0

Steps	$E(\ell)$	$E(\ell)$ ratio of MSE to model with $p = 3$, $d = 0$				
ℓ:	$p = 3$ $d = 0$	$p = 3$ $d = 1$	$p = 2$ $d = 0$	$p = 2$ $d = 1$	$p = 4$ $d = 0$	$p = 4$ $d = 1$
1	3.26	99.71	103.25	102.98	100.15	99.86
2	7.34	99.49	102.73	102.23	100.15	99.64
3	11.69	99.27	102.52	101.78	100.15	99.41
4	15.92	99.04	102.58	101.56	100.14	99.18
5	19.89	98.80	102.77	101.44	100.14	98.94
6	23.57	98.57	103.01	101.32	100.14	98.69
7	26.95	98.34	103.26	101.18	100.14	98.46
8	30.08	98.12	103.51	101.01	100.15	98.24
9	33.00	97.91	103.76	100.81	100.15	98.03
10	35.72	97.72	104.01	100.59	100.15	97.83
15	47.47	97.06	105.09	99.56	100.14	97.12
20	57.04	96.73	105.83	98.81	100.15	96.77
25	65.29	96.65	106.17	98.39	100.14	96.68
30	72.70	96.61	106.24	98.05	100.12	96.63
35	79.58	96.67	106.08	97.95	100.09	96.69
40	85.99	96.76	105.70	97.81	100.05	96.78
45	92.08	96.92	105.24	97.76	100.01	96.94
50	98.03	97.12	104.66	97.86	99.98	97.14
55	103.82	97.21	104.05	97.97	99.94	97.23
60	109.21	97.26	103.43	97.96	99.87	97.30

TABLE 4
Forecasting comparison for simulate model: true $p = 3$, roots are 0.5, 0.5, 0.99

Steps	$E(\ell)$	$E(\ell)$ ratio of MSE to model with $p = 3$, $d = 0$				
ℓ:	$p = 3$ $d = 0$	$p = 3$ $d = 1$	$p = 2$ $d = 0$	$p = 2$ $d = 1$	$p = 4$ $d = 0$	$p = 4$ $d = 1$
1	3.26	99.91	103.18	103.27	100.15	100.06
2	7.32	99.85	102.70	102.74	100.15	99.99
3	11.61	99.79	102.53	102.51	100.15	99.93
4	15.76	99.74	102.61	102.54	100.14	99.87
5	19.62	99.70	102.81	102.67	100.14	99.82
6	23.14	99.66	103.06	102.82	100.15	99.77
7	26.36	99.63	103.31	102.95	100.15	99.75
8	29.30	99.62	103.57	103.04	100.17	99.73
9	32.00	99.63	103.82	103.10	100.18	99.72
10	34.50	99.65	104.06	103.14	100.18	99.74
15	44.87	100.10	105.10	103.38	100.22	100.13
20	52.75	100.98	105.62	103.95	100.26	100.98
25	59.01	102.16	105.59	104.90	100.27	102.14
30	64.21	103.31	105.24	105.83	100.25	103.28
35	68.72	104.51	104.66	106.92	100.21	104.47
40	72.71	105.65	103.95	107.84	100.15	105.60
45	76.32	106.85	103.14	108.83	100.11	106.80
50	79.68	108.11	102.32	110.03	100.07	108.06
55	82.76	109.27	101.56	111.27	100.03	109.22
60	85.54	110.20	100.83	112.15	99.98	110.15

TABLE 5
Forecasting comparison for simulated model: true $p = 3$, roots are 0.5, 0.5, 0.95

Steps	$E(\ell)$	$E(\ell)$ ratio of MSE to model with $p = 3$, $d = 0$				
ℓ:	$p = 3$ $d = 0$	$p = 3$ $d = 1$	$p = 2$ $d = 0$	$p = 2$ $d = 1$	$p = 4$ $d = 0$	$p = 4$ $d = 1$
1	3.26	100.87	102.92	104.61	100.16	101.00
2	7.19	101.59	102.50	105.05	100.17	101.70
3	11.20	102.35	102.39	105.87	100.17	102.43
4	14.93	103.15	102.50	107.02	100.18	103.22
5	18.24	104.01	102.69	108.34	100.19	104.06
6	21.11	104.92	102.91	109.71	100.20	104.94
7	23.59	105.87	103.11	111.09	100.22	105.86
8	25.72	106.85	103.29	112.41	100.24	106.81
9	27.57	107.85	103.46	113.70	100.26	107.79
10	29.17	108.88	103.60	114.95	100.28	108.79
15	34.68	114.26	103.91	120.95	100.35	114.08
20	37.59	119.71	103.34	126.69	100.37	119.46
25	39.16	124.73	102.28	132.01	100.32	124.42
30	40.04	128.94	101.30	136.47	100.26	128.60
35	40.61	132.40	100.65	140.17	100.19	132.01
40	41.02	135.16	100.27	142.89	100.14	134.75
45	41.31	138.01	99.99	145.70	100.11	137.58
50	41.50	140.97	99.79	148.84	100.08	140.50
55	41.61	143.76	99.68	152.10	100.06	143.26
60	41.68	145.57	99.66	154.16	100.04	145.04

TABLE 6
Forecasting comparison for simulated data: true $p = 3$, roots are all 0.5

Steps	$E(\ell)$	$E(\ell)$ ratio of MSE to model with $p = 3$, $d = 0$				
ℓ:	$p = 3$ $d = 0$	$p = 3$ $d = 1$	$p = 2$ $d = 0$	$p = 2$ $d = 1$	$p = 4$ $d = 0$	$p = 4$ $d = 1$
1	3.26	105.23	100.68	108.65	100.13	104.82
2	5.90	110.01	100.58	114.55	100.13	109.11
3	7.70	115.20	100.65	121.43	100.13	113.78
4	8.74	120.67	100.72	128.63	100.13	118.68
5	9.28	126.00	100.77	135.33	100.14	123.50
6	9.53	130.73	100.78	140.98	100.14	127.93
7	9.64	134.52	100.74	145.38	100.14	131.62
8	9.68	137.31	100.64	148.61	100.13	134.39
9	9.69	139.30	100.49	150.91	100.12	136.32
10	9.69	140.71	100.32	152.55	100.10	137.64
15	9.68	144.26	99.98	156.66	100.04	140.87
20	9.68	145.69	99.98	158.33	100.02	142.22
25	9.66	147.22	99.98	160.24	100.00	143.62
30	9.66	148.04	99.98	161.31	100.00	144.37
35	9.67	148.70	99.98	162.52	100.00	144.86
40	9.68	148.98	99.99	162.92	100.00	145.11
45	9.68	150.37	99.99	164.35	100.00	146.48
50	9.69	153.41	99.98	167.67	99.99	149.35
55	9.68	155.18	99.99	169.99	99.99	150.98
60	9.66	155.33	99.99	170.78	99.99	150.92

only affects the forecast precision marginally. The $E(\ell)$ for all models are within 6% to the true model for all forecasts up to 60-step ahead.

To sum up, the simulation show that slight misspecification of AR order and under specification of unit root are not serious in forecasting but over-specification of unit root could result in poor forecast when the root of characteristic polynomial is far from 1. Yet, improvement of forecasting precision in absolute term could be substantial for large sample when the existence of unit root is appropriately taken into consideration.

6. Empirical results

6.1. Data

For empirical analysis, we analyze 6 most frequently used data sets in Taiwan including Gross Domestic Product (GDP), Consumer Price Indices (CPI), Wholesale Price Indices (WPI), Interest Rates(IR), Exchange Rate of New Taiwan Dollar to US Dollar(RX) and money supply(M1B). All series are quarterly data taken from the AREMOS databank. The sample period is 1961:1 to 1995:4 except for M1B which ranges between 1961:3 to 1995:4. So, sample size is 138 for M1B and 140 for the rest series. All series are seasonally unadjusted.

6.2. Order selection

Selecting lag order p and forecasting method simultaneously is analyzed in Ing, Lin and Yu [10]. Here, we follow the conventional wisdom by using AIC and chi-square statistics to determine p. When the AIC has a clear minimal, we select the order corresponding to the minimal AIC. When AIC is decreasing without a clear

minimum, we use chi-square statistics to select the last significant lag. It turns out that CPI, WPI and RX have order 2, interest rate has order 6, M1B has order 3 and GDP has order 8. The high order indicates the possible existence of seasonal unit root which is not investigated here.

6.3. Forecasting procedure

For each series, the first 100 observations are reserved for estimation and 1- to 20-step forecasts are computed. Then the model are re-estimated using first 101 observations and another 1- to 20-step forecasts are computed. The procedure is repeated until when the first $T-1$ observations are used to estimate the model and the last 1-step forecast is computed. Hence, we have 40 1-step forecasts, 39 2-step forecasts and 20 20-step forecasts except for M1B where there are 38 1-step forecasts and 18 20-step forecasts. For each step, the average root mean square error is computed.

6.4. Results

The results are reported in Tables 7 to 12. From the tables we observe the following. First, $E(\ell)$ increases linearly with ℓ for all series except for Interest Rates. This seems to suggest that except IR, all variables have a unit root. Second, regarding the Box-Jenkins forecast, imposing unit root constraint result in poor forecast for all steps ahead for WPI, CPI, GDP and IR. Especially for IR, the RMSE for $d=1$ is 200% higher than that for $d=0$. This seems to be consistent with the finding that its $E(\ell)$ converges to a constant very quickly. However, for RX forecast with $d=1$ fares much better than forecast with $d=0$. The precision gain from imposing unit root is about 5% for 1-step forecast and then up to over 30% for 20-step forecast. This seems to indirectly support the efficient market hypothesis for the foreign

TABLE 7
Forecasting comparison for GDP

		$E(\ell)$ ratio to model BJ, $d=0$		
ℓ	$E(\ell)$	BJ, $d=1$	Adap, $d=0$	Adap, $d=1$
1	12258.04	101.94	100.00	99.22
2	18024.46	104.69	101.47	179.52
3	21232.71	106.87	115.10	151.59
4	24719.28	108.70	106.82	83.76
5	31938.87	110.92	102.59	99.66
6	37316.81	112.05	116.19	125.65
7	40063.38	112.71	132.45	98.06
8	40505.82	109.83	147.52	45.57
9	46605.77	111.94	158.02	65.51
10	52966.58	116.89	159.53	103.16
11	57551.40	121.35	157.04	91.13
12	59480.32	120.18	169.46	66.06
13	66651.35	120.95	181.21	85.92
14	74760.98	123.52	184.47	109.41
15	79555.22	124.66	174.53	94.18
16	81162.26	120.64	173.44	74.78
17	91194.77	117.99	185.95	61.83
18	99975.45	122.09	181.84	99.95
19	105256.88	125.83	162.48	83.94
20	108809.60	122.63	162.14	55.65

TABLE 8
Forecasting comparison for CPI

steps		$E(\ell)$ ratio to model BJ, $d = 0$		
ℓ	$E(\ell)$	BJ, $d = 1$	Adap, $d = 0$	Adap, $d = 1$
1	1.12	99.39	100.00	101.46
2	1.71	98.81	83.53	77.07
3	1.82	99.20	100.10	69.33
4	1.83	99.75	123.42	79.90
5	2.09	98.49	128.11	84.27
6	2.51	99.73	124.42	81.85
7	2.54	101.83	148.32	89.80
8	2.65	102.66	165.72	95.17
9	2.97	102.13	163.20	95.48
10	3.19	101.81	171.06	95.60
11	3.06	106.53	209.48	98.15
12	3.12	108.45	237.77	96.83
13	3.55	106.61	236.77	86.76
14	3.71	107.33	257.23	89.74
15	3.76	111.25	289.99	98.66
16	3.83	113.36	326.41	99.02
17	4.32	110.28	336.41	91.27
18	4.50	111.09	372.91	89.71
19	4.44	113.86	431.77	86.81
20	4.79	111.22	455.83	80.87

TABLE 9
Forecasting comparison for WPI

		$E(\ell)$ ratio to model BJ, $d = 0$		
ℓ	$E(\ell)$	BJ, $d = 1$	Adap, $d = 0$	Adap, $d = 1$
1	1.16	102.47	100.00	101.19
2	2.13	103.43	58.78	93.97
3	3.05	104.54	50.13	97.35
4	3.88	105.55	47.53	96.59
5	4.56	107.43	49.39	108.12
6	5.07	109.64	53.84	113.24
7	5.41	112.38	61.27	113.87
8	5.58	116.54	69.88	120.83
9	5.89	120.00	76.71	132.83
10	6.36	121.66	81.36	140.82
11	6.93	122.71	86.87	137.69
12	7.53	123.95	90.96	135.62
13	8.02	125.92	96.49	141.24
14	8.48	127.93	103.98	150.77
15	8.82	130.43	113.97	157.73
16	8.87	135.06	127.60	157.82
17	8.97	139.02	141.47	173.71
18	9.15	142.11	157.87	191.82
19	9.52	143.65	174.20	193.95
20	10.19	142.43	186.53	178.22

TABLE 10
Forecasting comparison for RX

		$E(\ell)$ ratio to model BJ, $d = 0$		
ℓ	$E(\ell)$	BJ, $d = 1$	Adap, $d = 0$	Adap, $d = 1$
1	.66	95.38	100.00	99.40
2	1.32	92.38	53.40	90.39
3	2.01	89.53	40.91	81.93
4	2.77	88.16	37.25	75.67
5	3.41	87.09	38.60	73.91
6	3.99	86.26	42.45	72.13
7	4.42	85.01	47.25	72.20
8	4.71	83.30	54.59	70.22
9	5.03	82.30	61.77	68.81
10	5.29	81.93	68.35	72.37
11	5.58	81.90	74.77	75.55
12	5.94	81.98	79.20	77.42
13	6.31	81.57	81.79	76.30
14	6.67	80.44	83.67	74.67
15	6.93	78.33	85.17	74.42
16	7.11	75.67	86.40	70.66
17	7.34	72.78	86.34	62.28
18	7.56	70.48	86.23	58.96
19	7.87	69.58	85.50	59.46
20	8.24	69.51	83.95	51.02

TABLE 11
Forecasting comparison for M1B

		$E(\ell)$ ratio to model BJ, $d = 0$		
ℓ	$E(\ell)$	BJ, $d = 1$	Adap, $d = 0$	Adap, $d = 1$
1	96297.39	92.81	100.00	102.13
2	152389.01	94.00	65.10	86.82
3	208305.92	94.22	52.25	78.01
4	266876.68	94.73	52.56	71.28
5	377584.71	92.59	55.08	61.12
6	481271.98	94.39	61.18	56.24
7	532886.35	99.13	59.46	55.99
8	595646.12	101.27	71.45	49.52
9	668073.88	108.19	84.69	50.29
10	774390.51	113.50	92.29	49.31
11	821482.40	123.40	89.06	50.77
12	886619.83	129.37	90.62	46.46
13	1052170.67	129.01	89.81	42.85
14	1158059.45	143.04	86.93	44.22
15	1335812.44	145.93	70.46	44.98
16	1378939.42	165.55	60.58	42.82
17	1649465.49	162.79	56.17	46.78
18	1748042.18	183.13	61.66	45.31
19	1893323.08	200.50	51.17	41.98
20	2079603.50	214.50	50.45	28.34

TABLE 12
Forecasting comparison for IR

		$E(\ell)$ ratio to model BJ, $d = 0$		
ℓ	$E(\ell)$	BJ, $d = 1$	Adap, $d = 0$	Adap, $d = 1$
1	0.72	104.58	100.00	106.60
2	1.16	108.68	67.94	79.35
3	1.24	114.84	64.70	108.65
4	1.38	120.28	58.85	127.00
5	1.63	124.48	48.70	125.50
6	1.82	130.14	46.05	120.60
7	1.83	138.25	50.44	124.52
8	1.82	146.64	55.74	139.96
9	1.83	154.97	58.67	170.56
10	1.83	164.17	59.64	191.38
11	1.76	175.17	60.49	208.55
12	1.70	185.17	63.08	211.01
13	1.67	194.23	65.27	208.26
14	1.61	204.93	75.07	213.25
15	1.52	216.39	91.16	219.55
16	1.36	237.94	94.70	238.53
17	1.27	254.73	108.89	239.50
18	1.10	289.28	101.44	237.00
19	0.84	370.74	131.31	269.90
20	0.59	521.61	199.32	349.14

exchange market in Taiwan. As for M1B, imposing unit root constraint improves forecast precision from 1- to 7-step forecasts but deteriorates forecast precision from 8-step to 20-step forecasts. The inefficiency is more than 100% for 19 and 20-step forecasts. Third, the performance of adaptive forecaster is mixed. For RX and M1B, adaptive forecast with $d = 0$ and $d = 1$ consistently outperforms conventional Box-Jenkins' forecast by a large margin. The precision gain could go as high as 50%. For CPI adaptive forecast performs poorly for $d = 0$ but very well for $d = 1$. For IR and WPI adaptive forecast with $d = 0$ performs well in short and medium term forecast but fares poorly in long term forecast. But adaptive forecast with $d = 1$ performs okay in the short term but very poorly in the long term. The case GDP is quite interesting. While adaptive forecast with $d = 0$ fares poorly for short and long term forecast, the performance of adaptive forecast with $d = 1$ jumps up and down across steps. This seems to suggest that seasonality plays an important for the differenced GDP which is supported by the corresponding autocorrelation function. This issue will be investigated in future study.

To sum up, the empirical findings are mixed. Imposing unit root constraint might improve forecast precision for some cases but deteriorate forecast precision in others. Also, adaptive forecast differs from Box-Jenkins' forecast by the big margin. Most frequently, it could improve short to medium term forecast but result in poor long term forecast. However, for some cases, it could produce either better or worse forecast for forecast of all steps. Further study is needed to determine the influencing factors.

7. Conclusions

We have analyzed the least square forecaster from various aspects. From the theoretical viewpoint, we prove that C_T, the most important quantity when evaluating the performance of 1-step forecasters is equal to $(p + d)\sigma^2 \log(T)$ where d is 1 or 0

depending if there is a unit root. This result could be used to analyze the gain in forecasting precision when unit root is detected and is taken into account. Further, this theorem can lead to a simple proof of the strong consistency of PLS in AR model selection and a new test of unit root.

Our simulation analysis confirms the theoretical results. In addition, we also learn that while mis-specification of AR order has marginal impact on forecasting precision over-specification of unit root strongly deteriorate the quality of long term forecast. As for the empirical study using Taiwanese data, the result is mixed. Adaptive forecast and imposing unit root improves forecast precision for some cases but deteriorates forecasting precision for other cases.

Acknowledgments

Financial support from the National Science Council under grant NSC85-2415-H001-009 is gratefully acknowledged. We would like to thank C. K. Ing and two anonymous referees for helpful comments and suggestions. Without mentioning, the authors are responsible for any remaining error.

References

[1] Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics* **48** 577–602.

[2] Bhansali, R. J. (1997). Direct autoregressive predictors for multistep prediction: order selection and performance relative to the plug in predictors. *Statistica Sinica* **7** 425–449.

[3] Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *The Annals of Statistics* **16** 367–401.

[4] Chow, Y. S. (1965). Local convergence of martingales and the law of large numbers. *The Annals of Mathematical Statistics* **36** 552–558.

[5] Eberlein, E. and Taqqu, M. S. (1986). *Dependence in Probability and Statistics*. Birkhäuser, Boston.

[6] Goodwin G. C. and Sin, K. S. (1984). *Adaptive Filtering, Prediction and Control*. Prentice-Hall, Englewood Cliffs, NJ.

[7] Ing, C. K. (2001). A note on mean-squared prediction errors of the least squares predictors in random walk models. *Journal of Time Series Analysis* **22** 711–724.

[8] Ing, C. K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory* **19** 254–279.

[9] Ing, C. K. (2004). Selecting optimal multistep predictors for autoregressive processes of unknown order. *The Annals of Statistics* **32** 693–722.

[10] Ing, C. K., Lin, J. L. and Yu, S. H. (2006). Toward optimal multistep forecasts in unstable autoregressions. Manuscript.

[11] Ing, C. K. and Wei, C. Z. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics* **33** 2423–2474.

[12] Lai, T. L. and Wei, C. Z. (1982). Asymptotic properties of projections with applications to stochastic regression problems. *Journal of Multivariate Analysis* **12** 346–370.

[13] Lin and Tsay (1996). Cointegration constraint and forecasting: An empirical examination. *Journal of Applied Econometrics* **11** 519–538.
[14] Rissanen, J. (1986). Order estimation by accumulated prediction errors. In *Essays in Time Series and Applied Processes*, Special vol. 23A of *Journal of Applied Probability* 55–61.
[15] Tiao, G. C. and Tsay, R. S. (1994). Some advances in nonlinear and adaptive modelling in time series. *Journal of Forecasting* **13** 109–131.
[16] Tsay, R. S. (1983). Order selection under nonstationary autoregressive and stochastic regression models. *The Annals of Statistics* **12** 1425–1433.
[17] Wei, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with application to time series, *Annals of Statistics* **15** 1667–1682.
[18] Wei, C. Z. (1992). On predictive least square principles. *The Annals of Statistics* **20** 1–42.
[19] Weiss, A. (1991). Multi-step estimation and forecasting in dynamic models. *Journal of Econometrics* **48** 135–149.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 93–112

DOI: 10.1214/074921706000000978

Order determination in general vector autoregressions

Bent Nielsen[1]

University of Oxford

Abstract: In the application of autoregressive models the order of the model is often estimated using either a sequence of likelihood ratio tests, a likelihood based information criterion, or a residual based test. The properties of such procedures has been discussed extensively under the assumption that the characteristic roots of the autoregression are stationary. While non-stationary situations have also been considered the results in the literature depend on conditions to the characteristic roots. It is here shown that these methods for lag length determination can be used regardless of the assumption to the characteristic roots and also in the presence of deterministic terms. The proofs are based on methods developed by C. Z. Wei in his joint work with T. L. Lai.

1. Introduction

Order determination for stationary autoregressive time series has been discussed extensively in the literature. The three prevailing methods are either to test redundance of the last lag using a likelihood based test, to estimate the lag length consistently using an information criteria, or to investigate the residuals of a fitted model with respect to autocorrelation. It is shown that these methods can be used regardless of any assumptions to the characteristic roots. This is important in applications, as the question of lag length can be addressed without having to locate the characteristic roots.

The statistical model is given by a p-dimensional time series X_t of length $K+T$ satisfying a Kth order vector autoregressive equation

$$X_t = \sum_{l=1}^{K} A_l X_{t-l} + \mu D_t + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1.1}$$

conditional on the initial values $X_0, \ldots, X_{1-K}$. The effective sample will remain $X_1, \ldots, X_T$ when discussing autoregressions with $k < K$ to allow comparison of likelihood values. The component D_t is a vector of deterministic terms such as a constant, a linear trend, or seasonal dummies. For the sake of defining a likelihood function it is initially assumed that the innovations, (ε_t), are independently, identically normal, $\mathsf{N}_p(0, \Omega)$, distributed and independent of the initial values.

The aim is to determine the largest non-trivial order for the time series, k_0 say with $0 \leq k_0 \leq K$, so $A_{k_0} \neq 0$ and $A_j = 0$ for $j > k_0$. Three approaches are available of which the first is based on a likelihood ratio test for $A_k = 0$ where $1 \leq k \leq K$. The log likelihood ratio test statistic is

$$\mathsf{LR}\,(A_k = 0) = T \log\det \hat{\Omega}_{k-1} - T \log\det \hat{\Omega}_k,$$

[1]Department of Economics, University of Oxford & Nuffield College, Oxford OX1 1NF, UK, e-mail: bent.nielsen@nuffield.ox.ac.uk

AMS 2000 subject classifications: primary 62M10; secondary 62F10.
Keywords and phrases: autoregression, lag length, information criteria.

where $\hat{\Omega}_{k-j}$ is the conditional maximum likelihood estimator based on the observations $X_1, \ldots, X_T$ given the initial values, see (3.2) below. The statistic LR is proved to be asymptotically χ^2 under the hypothesis $k_0 < k$, generalising results for the purely non-explosive case. Since the result does not depend on the characteristic roots, it can be used for lag length determination before locating the characteristic roots.

The second approach is to estimate k_0 by the argument $\hat{k}$ that maximises a penalised likelihood, or equivalently, minimises an information criteria of the type

$$\Phi_j = \log\det\hat{\Omega}_j + j\frac{f(T)}{T}, \qquad j = 0, \ldots, K. \tag{1.2}$$

In the literature there are several candidates for the penalty function f. Akaike has $f(T) = 2p^2$, Schwarz [23] has $f(T) = p^2 \log T$ while Hannan and Quinn [10] and Quinn [22] have $f(T) = 2p^2 \log\log T$. For stationary processes without deterministic components it has been shown that the estimator $\hat{k}$ is weakly consistent if $f(T) = \mathrm{o}(T)$ and $f(T) \to \infty$ as T increases, while Hannan and Quinn show, for $p = 1$, that strong consistency is obtained if $f(T) = \mathrm{o}(T)$ and $\liminf_{T\to\infty} f(T)/\log\log T > 2$, while strong consistency cannot be obtained if $\limsup_{T\to\infty} f(T)/\log\log T < 2$. In other words the estimators of Hannan and Quinn and of Schwarz are consistent while Akaike's estimator is inconsistent. Some generalisations to non-explosive processes have been given by for instance Paulsen [20], Pötscher [21] and Tsay [24]. Pötscher also considered the purely explosive case but did not obtain a common feasible rate for $f(T)$ for the explosive and the non-explosive case. In the following consistency is shown for a penalty function $f(T)$ not depending on the characteristic roots, showing that the penalised likelihood approach also can be applied to lag length determination prior to locating the characteristic roots.

A third approach is a residual based mis-specification test. This is implemented in particular in econometric computer packages. In a first step the residuals, $\hat{\varepsilon}_t$ say, are computed from the model (1.1) with $k-1$ lags, say. In a second step an auxillary regression is considered where $\hat{\varepsilon}_t$ is regressed on lagged values as well as the regressors in equation (1.1). It is argued that a test based on the squared multiple correlation arising from the auxillary regression is asymptotically equivalent to the above mentioned likelihood ratio test statistic also in the general case.

Like the work of Pötscher [21] the proofs in this paper are based on the joint work of C. Z. Wei and T. L. Lai on the strong consistency of least squares estimators in autoregressions, see for instance Lai and Wei [15]. As pointed out in Pötscher's Remark 1 to his Theorem 3.3 these results are not quite strong enough to facilitate common feasible rates for the penalty function. Two important ingredients in the presented proofs are therefore an algebraic decomposition exploiting partitioned inversion along with a generalisation of Lai and Wei's work given by Nielsen [17]. Whereas the former paper is concerned with showing that the least squares estimator for the autoregressive estimator is consistent, the latter paper provides a more detailed discussion of the rate of consistency as well as it allows deterministic terms in the autoregression.

The following notation is used throughout the paper: For a quadratic matrix α let $\mathrm{tr}(\alpha)$ denote the trace and $\lambda(\alpha)$ the set of eigenvalues, so that $|\lambda(\alpha)| < 1$ means that all eigenvalues have absolute value less than one. When α is also symmetric then $\lambda_{\min}(\alpha)$ and $\lambda_{\max}(\alpha)$ denote the smallest and the largest eigenvalue respectively. The abbreviations *a.s.* and P are used for properties holding almost surely and in probability, respectively.

2. Results

Before presenting the results the assumptions and notation is set up. Then the results follow for the three approaches.

2.1. Assumptions and notation

The asymptotic analysis is to a large extent based on results of Lai and Wei [15] with appropriate modifications to the situation with deterministic terms in Nielsen [17]. Following that analysis the assumption to the innovations of independence and normality made above can be relaxed so that the sequence of innovations (ε_t) is a martingale difference sequence with respect to an increasing sequence of σ-fields $(\mathcal{F}_t)$, that is: the innovations $X_{1-k}, \ldots, X_0$ are $\mathcal{F}_0$-measurable and ε_t is $\mathcal{F}_t$-measurable with $\mathsf{E}(\varepsilon_t|\mathcal{F}_{t-1}) \overset{a.s.}{=} 0$, which is assumed to satisfy

$$\sup_t \mathsf{E}\{(\varepsilon_t'\varepsilon_t)^{\lambda/2}|\mathcal{F}_{t-1}\} \overset{a.s.}{<} \infty \qquad \text{for some } \lambda > 4. \tag{2.1}$$

To establish an asymptotic theory for the LR-statistic it is assumed that

$$\mathsf{E}\left(\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\right) \overset{a.s.}{=} \Omega, \tag{2.2}$$

where Ω is positive definite. For the asymptotic theory for the information criteria this can be relaxed to

$$\liminf_{t\to\infty} \lambda_{\min} \mathsf{E}\left(\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}\right) \overset{a.s.}{>} 0. \tag{2.3}$$

The deterministic term D_t is a vector of terms such as a constant, a linear trend, or periodic functions like seasonal dummies. Inspired by Johansen [13] the deterministic terms are required to satisfy the difference equation

$$D_t = \mathbf{D}D_{t-1}, \tag{2.4}$$

where $\mathbf{D}$ has characteristic roots on the complex unit circle. For example,

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix} \qquad \text{with} \qquad D_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

will generate a constant and a dummy for a biannual frequency. The deterministic term D_t is assumed to have linearly independent coordinates. That is:

$$|\lambda(\mathbf{D})| = 1, \qquad \operatorname{rank}(D_1, \ldots, D_{\dim \mathbf{D}}) = \dim \mathbf{D}. \tag{2.5}$$

In the analysis it is convenient to introduce the companion form

$$\begin{pmatrix} \mathbf{X}_t \\ D_t \end{pmatrix} = \begin{pmatrix} \mathbf{B} & \boldsymbol{\mu} \\ 0 & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{t-1} \\ D_{t-1} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \end{pmatrix}, \tag{2.6}$$

where $\mathbf{X}_{t-1} = (X_{t-1}', \ldots, X_{t-k+1}')'$ and

$$\mathbf{B} = \begin{Bmatrix} A_1 \cdots A_{k-2} & A_{k-1} \\ I_{p(k-2)} & 0 \end{Bmatrix}, \quad \boldsymbol{\iota} = \begin{Bmatrix} I_p \\ 0_{(k-2)p\times p} \end{Bmatrix}, \quad \boldsymbol{\mu} = \boldsymbol{\iota}\mu\mathbf{D}, \quad \mathbf{e}_t = \boldsymbol{\iota}\varepsilon_t.$$

The process $\mathbf{X}_t$ can be decomposed using a similarity transformation. Following Herstein ([11], p. 308) there exists a regular, real matrix M that block-diagonalises

$\mathbf{B}$ so that $M\mathbf{B}M^{-1} = \text{diag}(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is a real block diagonal matrix where the eigenvalues of the diagonal blocks $\mathbf{U}, \mathbf{V}, \mathbf{W}$ satisfy $|\lambda(\mathbf{U})| < 1$, $|\lambda(\mathbf{V})| = 1$, and $|\lambda(\mathbf{W})| > 1$. Any of the blocks $\mathbf{U}, \mathbf{V}, \mathbf{W}$ can be empty matrices, so if for instance $|\lambda(\mathbf{B})| < 1$ then $\mathbf{U} = \mathbf{B}$ and $\dim \mathbf{V} = \dim \mathbf{W} = 0$. The process $\mathbf{X}_t$ can therefore be decomposed as

$$M\mathbf{X}_t = \begin{pmatrix} U_t \\ V_t \\ W_t \end{pmatrix} = \begin{pmatrix} \mathbf{U} & 0 & 0 & \mu_U \\ 0 & \mathbf{V} & 0 & \mu_V \\ 0 & 0 & \mathbf{W} & \mu_W \end{pmatrix} \begin{pmatrix} U_{t-1} \\ V_{t-1} \\ W_{t-1} \\ D_t \end{pmatrix} + \begin{pmatrix} e_{U,t} \\ e_{V,t} \\ e_{W,t} \end{pmatrix}. \tag{2.7}$$

Finally, there exists a constant $\tilde{\mu}_U$, see Nielsen ([17], Lemma 2.1), so

$$U_t = \tilde{U}_t + \tilde{\mu}_U D_t \qquad \text{where} \qquad \tilde{U}_t = \mathbf{U}\tilde{U}_{t-1} + e_{U,t}. \tag{2.8}$$

2.2. Likelihood ratio test statistics

The likelihood ratio test statistic is known to be asymptotically χ^2 in the stationary case where $|\lambda(\mathbf{B})| < 1$ and $\mathbf{D} = 1$, see Lütkepohl ([16], Section 4.2.2). Here the result is shown to hold regardless of the assumptions to $\mathbf{B}$ and $\mathbf{D}$. Thus, the likelihood ratio test can be used before locating the charateristic roots.

Theorem 2.1. *Suppose Assumptions* (2.1), (2.2), (2.5) *are satisfied and* $k_0 < k$. *Then* $\mathsf{LR}(A_k = 0)$ *is asymptotically* $\chi^2(p^2)$.

Since the likelihood ratio test statistic is based on partial correlations it follows from Theorem 2.1 that partial correlograms that are computed from partial correlograms can be used regardless of the location of the characteristic roots. Often correlograms are, however, based on the Yule-Walker estimators, which assume stationarity. For non-stationary autoregressions that can lead to misleading inference. Nielsen [18] provides a more detailed discussion.

Remark 2.2. The fourth order moment condition, $\lambda > 4$, in Assumption (2.1) is used twice in the proof. First, to ensure that the residuals from regressing ε_t on the explosive term W_{t-1} do not depend asymptotically on W_{t-1}. As discussed in Remark 3.7 it suffices that $\lambda > 2$ if either of the following conditions hold:

(I,a) $\dim \mathbf{W} = 0$.
(I,b) $\dim \mathbf{W} > 0$ and ε_t independent, identically distributed.

Secondly, to ensure that $\varepsilon_t \varepsilon_{t-1}$ has second moments when applying a Central Limit Theorem. As discussed in Remark 3.12, it suffice that $\lambda > 2$ if

(II) the innovations ε_t are independent.

The test statistic considered above is for a hypothesis concerning a single lag. This can be generalised to a hypothesis concerning several lags, m say, where $k + m - 1 \leq K$.

Theorem 2.3. *Suppose Assumptions* (2.1), (2.2), (2.5) *are satisfied and* $k_0 < k$. *Then* $\mathsf{LR}(A_k = \cdots = A_{k+m-1} = 0)$ *is asymptotically* $\chi^2(p^2 m)$.

2.3. Information criteria

The next two results concern consistency of a lag length estimator arising from use of information criterions. The proof has two distinct parts. First, it is argued that

the lag length estimator $\hat{k}$ is not under-estimating, and, secondly, that it is not over-estimating. The first part is the easy one to establish. This result holds for all of the penalty functions discussed in the introduction under weak conditions to the innovations.

Theorem 2.4. *Suppose Assumptions* (2.1), (2.3), (2.5) *are satisfied with* $\lambda > 2$ *only and* $f(T) = \mathrm{o}(T)$. *Then* $\liminf_{T\to\infty} \hat{k} \overset{a.s.}{\geq} k_0$.

This result has previously been established in the univariate case without deterministic terms so $p = \dim X = 1$ and $\dim \mathbf{D} = 0$ by Pötscher (1989, Theorem 3.3). For the purely explosive case $|\lambda(\mathbf{B})| > 1$ his Theorem 3.2 shows the above result under the weaker condition $f(T) = \mathrm{o}(T^2)$. A version holding in probabilty has been shown for the non-explosive case $|\lambda(\mathbf{B})| \leq 1$ and $\mathbf{D} = 1$ by Paulsen [20] and Tsay [24].

Results showing that the lag length is not overestimating are harder to establish. Various weak and strong results can be obtained depending on the number of conditions that are imposed.

Theorem 2.5. *Suppose Assumptions* (2.1), (2.5) *are satisfied. Then*

(i) *If* $f(T) \to \infty$ *and Assumption* (2.2) *holds then* $\mathsf{P}(\hat{k} \leq k_0) \to 1$.

(ii) *If* $f(T)/\log T \to \infty$ *and Assumption* (2.3) *holds then* $\limsup_{T\to\infty} \hat{k} \overset{a.s.}{\leq} k_0$.

(iii) *If* $f(T)/\{(\log\log T)^{1/2}(\log T)^{1/2}\} \to \infty$, *Assumption* (2.3) *holds, and the parameters satisfy the condition* (A) *that* $\mathbf{V}$ *and* $\mathbf{D}$ *have no common eigenvalues then* $\limsup_{T\to\infty} \hat{k} \overset{a.s.}{\leq} k_0$.

(iv) *If* $f(T)/\log\log T \to \infty$, *Assumption* (2.3) *holds, and either* $(B)\dim \mathbf{D} = 0$ *with* $\mathbf{V} = 1$ *or* $(C)\dim \mathbf{V} = 0$ *then* $\limsup_{T\to\infty} \hat{k} \overset{a.s.}{\leq} k_0$.

(v) *Suppose Assumption* (2.2) *holds, and either* (B) *or* (C) *holds then*

(a) *If* $\liminf_{T\to\infty}(2\log\log T)^{-1} f(T) \overset{a.s.}{>} p^2$ *then* $\limsup_{T\to\infty} \hat{k} \overset{a.s.}{\leq} k_0$.

(b) *If* $\limsup_{T\to\infty}(2\log\log T)^{-1} f(T) \overset{a.s.}{<} 1$ *then* $\hat{k} \overset{a.s.}{\nrightarrow} k_0$.

By combining Theorems 2.4, 2.5 consistency results can be obtained. For instance Theorem 2.4 in combination with Theorem 2.5(i) shows that $\hat{k} \xrightarrow{\mathsf{P}} k_0$ if the penalty function satisfies $f(T) \to \infty$ and $f(T) = \mathrm{o}(T)$. This includes Hannan and Quinn's and Schwarz's penalty functions, but excludes that of Akaike as usually found. Likewise, Theorem 2.4 in combination with Theorem 2.5(ii) show that $\hat{k} \overset{a.s.}{\to} k_0$ if the penalty function satisfies $f(T)/\log T \to \infty$ and $f(T) = \mathrm{o}(T)$. These results are the first to present conditions to the penalty function ensuring consistency that are not depending on the parameter $\mathbf{B}$ and $\mathbf{D}$. This implies that the information criteria can be used before locating the charateristic roots.

It remains an open problem, however, to establish strong consistency of the Schwarz and the Hannan-Quinn estimators for general values of $\mathbf{V}$ and $\mathbf{D}$. Theorem 2.4 combined with Theorem 2.5(iii) shows that the Schwarz estimator is strongly consistent when (A) holds so $\mathbf{V}$ and $\mathbf{D}$ have no common eigenvalues. Theorem 2.4 combined with Theorem 2.5(v) shows that the Hannan-Quinn estimator is strongly consistent when either (B) $\dim \mathbf{D} = 0$ with $\mathbf{V} = 1$ or $(C)\dim \mathbf{V} = 0$ holds. This is the first strong consistency result for the Hannan-Quinn estimator in the non-stationary case.

Remark 2.6. In Theorem 2.5 the fourth order moment condition $\lambda > 4$ in Assumption (2.1) can be relaxed to $\lambda > 2$ under certain condions to the parameters. Recall the conditions stated in Remark 2.2 which are

(I,a) $\dim \mathbf{W} = 0$.
(I,b) $\dim \mathbf{W} > 0$ and ε_t independent, identically distributed.
(II) the innovations ε_t are independent.

As discussed in Remark 3.13 it holds:
Result (i) can be relaxed if (II) holds along with either (I,a) or (I,b).
Results (ii), (iii), (iv) can be relaxed if (I,a) holds.
Result (v) cannot be relaxed with the present proof.

A number of related results are available in the literature.

The weak consistency results in (i) has been shown for the non-explosive case $|\lambda(\mathbf{B})| \leq 1$ and $\mathbf{D} = 1$ by Paulsen [20] and Tsay [24].

The $(\log\log T)^{1/2}(\log T)^{1/2}$ rate discussed in Theorem 2.5(iii) and Remark 2.6(iii) is an improvement over the $\log T$ rates discussed by for instance Pötscher [21] and Wei [25]. These authors discuss the univariate case without deterministic terms so $p = \dim X = 1$ and $\dim \mathbf{D} = 0$, in which case $\mathbf{V}$ and $\mathbf{D}$ trivially have no common eigenvalues. *First*, Pötscher ([21], Theorem 3.1) shows an under-estimation result for rates satisfying $f(T)/\log T \to \infty$ in the non-explosive case so $|\lambda(\mathbf{B})| \leq 1$, hence $\dim \mathbf{W} = 0$, but with Assumption (2.3) replaced by the weaker condition that $\liminf_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathsf{E}(\varepsilon_t^2|\mathcal{F}_{t-1}) \overset{a.s.}{\to} 0$. Pötcher's Theorem 3.2 concerning under-estimation in the purely explosive case so $|\lambda(\mathbf{B})| > 1$ requires $\liminf_{T\to\infty} f(T)/T > 0$ *a.s.* with just $\lambda > 2$ in Assumption (2.1). The Remark 1 to his Theorem 3.3 points out that his results do not provide a common feasibility rate for autoregressions with both explosive and non-explosive roots in that $f(T) = \mathrm{o}(T)$ is required for the over-estimation result, whereas $\liminf_{T\to\infty} f(T)/T > 0$ *a.s.* is required for the under-estimation results. *Secondly*, Theorem 3.6 of Wei [25] goes a step further in showing the over-estimation result for the rate $f(T) = \log T$ for the non-explosive case so $\dim \mathbf{W} = 0$.

The optimal $\log\log T$ rates in (v) were originally suggested by Hannan and Quinn [10] and Quinn [22] for the case where $|\lambda(\mathbf{B})| < 1$, $\dim \mathbf{D} = 0$. A full generalisation cannot be made at present as the proof hinges on proving that the smallest eigenvalue of the average of the squared residual from regressing V_{t-1} on D_t, that is $T^{-1-\eta}\sum_{t=1}^{T}(V_{t-1}|D_t)(V_{t-1}|D_t)'$, has positive limit points for some $\eta > 0$. This result can only be established in two special cases: first, if $\dim \mathbf{V} = 0$ the issue is irrelevant, and secondly, if $\mathbf{V} = 1$ and $\dim \mathbf{D} = 0$ this follows from the law of iterated logarithms by Donsker and Varadhan [6]. A more detailed discussion is given in Lemma 3.5(iv) in the Appendix.

The strong $\log\log T$ rate in Theorem 2.5(iv) and Remark 2.6(iv) has previously been established in the purely stable, univariate case without deterministic terms, so $p = \dim X = 1$ and $\dim \mathbf{D} = 0$ and $|\lambda(\mathbf{B})| < 1$, and hence $\dim \mathbf{W} = 0$, see Pötscher ([21], Theorem 3.4). Once again, his result only requires $\liminf_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathsf{E}(\varepsilon_t^2|\mathcal{F}_{t-1}) \to 0$ *a.s.* instead of Assumption 2.3.

2.4. Residual based mis-specification testing

The third approach is to fit the model (1.1) with $k-1$ lags and analyse the residuals for autocorrelation of order up to m. The maximal lag length parameter K is here required to be at least $k-1$. This is done in two steps. First the residuals $\hat{\varepsilon}_t$ are

found for the regression (1.1) with $t = 1, \ldots, T$ and $k-1$ lags. In the second step $\hat{\varepsilon}_t$ is analysed in an auxillary regression for $t = m+1, \ldots, T$, where $\hat{\varepsilon}_t$ is regressed on $\hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-m}$ as well as the original regressors $\mathbf{X}_{t-1} = (X'_{t-1}, \ldots, X'_{t-k+1})'$ and D_t. The original regressors are included to mimic the above likelihood analysis where $\mathbf{X}_{t-1}, D_t$ are partialled out from X_t and X_{t-k}. A test based on the squared sample correlation of the variables in the auxillary regression is asymptotically equivalent to the likelihood ratio tests, so the degrees of freedom do not include the dimension of $\mathbf{X}_{t-1}, D_t$. In the multivariate case, $p > 1$, the test can be implemented in three ways, using either a simultaneous test, a marginal test or a conditional test.

The joint test, is based on the test statistic $\mathrm{tr}(TR^2)$, where R^2 is the squared sample multiple correlation of $\hat{\varepsilon}_t$ and $(\hat{\varepsilon}'_{t-1}, \ldots, \hat{\varepsilon}'_{t-m}, \mathbf{X}'_{t-1}, D'_t)'$.

The other two tests are based on a q-dimensional subset of the p components of ε_t. As the equations in the model equation (1.1) can be permuted there is no loss of generality in focussing on the first q components. Thus, partition

$$\varepsilon_t = \begin{pmatrix} \varepsilon_{t,1} \\ \varepsilon_{t,2} \end{pmatrix}, \qquad X_t = \begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix},$$

where $\varepsilon_{t,1}$ and $X_{t,1}$ are q-dimensional.

The marginal model consists of the first q equations of (1.1), that is $X_{t,1}$ given $\mathbf{X}_{t-1}, D_t$. The marginal test is then based on the squared sample multiple correlation, R^2_{marg} say, of $\hat{\varepsilon}_{t,1}$ and $(\hat{\varepsilon}'_{t-1,1}, \ldots, \hat{\varepsilon}'_{t-m,1}, \mathbf{X}'_{t-1}, D'_t)$.

The conditional model consists of the first q equations of (1.1) given $X_{t,2}$, that is $X_{t,1}$ given $X_{t,2}, \mathbf{X}_{t-1}, D_t$. The conditional test is based on the squared sample multiple correlation, R^2_{cond} say, of $\hat{\varepsilon}_{t,1}$ and $(\hat{\varepsilon}'_{t-1,1}, \ldots, \hat{\varepsilon}'_{t-m,1}, X'_{t,2}, \mathbf{X}'_{t-1}, D'_t)$.

The following asymptotic result can be established.

Theorem 2.7. *Suppose Assumptions* (2.1), (2.2), (2.5) *are satisfied and* $k_0 < k$. *Then* $\mathrm{tr}(TR^2)$ *is asymptotically* $\chi^2(p^2m)$, *while* $\mathrm{tr}(TR^2_{marg})$ *and* $\mathrm{tr}(TR^2_{cond})$ *are asymptotically* $\chi^2(q^2m)$.

Sometimes these test are implemented so that the auxillary regression is carried out for $t = 1, \ldots, T$ rather than $t = m+1, \ldots, T$ with the convention that $\hat{\varepsilon}_0 = \cdots = \hat{\varepsilon}_{1-m} = 0$. Variants of the tests have been considered, in particular for the univariate case, by Durbin [7], Godfrey [8], Breusch [3] and Pagan [19]. Those variants have been argued to be score/Lagrange multiplier type tests and asymptotic theory has been established for the stationary case $|\lambda(\mathbf{B})| < 1$.

3. Proofs

The likelihood ratio test statistic for testing $A_k = 0$ is given by

$$\begin{aligned} LR\,(A_k = 0) &= -T \log\det(\hat{\Omega}_{k-1}^{-1}\hat{\Omega}_k) \\ &= -T \log\det\{I_p - \hat{\Omega}_{k-1}^{-1}(\hat{\Omega}_{k-1} - \hat{\Omega}_k)\}, \end{aligned} \tag{3.1}$$

where $\hat{\Omega}_k$ and $\hat{\Omega}_{k-1}$ represent the unrestricted and restricted maximum likelihood estimators for the variance matrix defined below. In the following first some notation is introduced. Then comes an asymptotic analysis of $\hat{\Omega}_{k-1}$ and $\hat{\Omega}_{k-1} - \hat{\Omega}_k$ and finally proofs of the main theorems follow.

3.1. Notation

It is convenient to introduce some notation to handle $\hat{\Omega}_{k-1}$ as well as $\hat{\Omega}_{k-1} - \hat{\Omega}_k$. Thus, let the residuals from the partial regressions of X_t and X_{t-k} on $\mathbf{X}_{t-1} = (X'_{t-1}, \ldots, X'_{t-k+1})'$ and the deterministic components D_t be denoted

$$(X_t|\mathbf{X}_{t-1}, D_t), \qquad (X_{t-k}|\mathbf{X}_{t-1}, D_t).$$

When the hypothesis, $A_k = 0$, is satisfied then $(X_t|\mathbf{X}_{t-1}, D_t) = (\varepsilon_t|\mathbf{X}_{t-1}, D_t)$ and therefore the restricted variance estimator is given by

$$\hat{\Omega}_{k-1} = \frac{1}{T}\sum_{t=1}^{T} (\varepsilon_t|\mathbf{X}_{t-1}, D_t)\,(\varepsilon_t|\mathbf{X}_{t-1}, D_t)'. \tag{3.2}$$

Most of the analysis in the proof relates to $\hat{\Omega}_{k-1} - \hat{\Omega}_k$ so it is helpful to define

$$Q(Z_t) = \sum_{t=1}^{T} \varepsilon_t Z'_t \left(\sum_{t=1}^{T} Z_t Z'_t\right)^{-1} \sum_{t=1}^{T} Z_t \varepsilon'_t,$$

for any time series Z_t. It follows that $T(\hat{\Omega}_{k-1} - \hat{\Omega}_k) = Q(X_{t-k}|\mathbf{X}_{t-1}, D_t)$. Occasionally the following notation will be used: For a matrix α let $\alpha^{\otimes 2} = \alpha\alpha'$.

3.2. Asymptotic analysis of $\hat{\Omega}_{k-1}$

Asymptotic expressions for the restricted least squares variance estimator $\hat{\Omega}_{k-1}$ are given by Nielsen ([17], Corollary 2.6, Theorem 2.8):

Lemma 3.1. *Suppose $A_k = 0$ and that the Assumptions (2.1), (2.3), (2.5) are satisfied with $\lambda > 2$. Then, for all $\xi < 1 - 2/\lambda$ it holds*

$$\hat{\Omega}_{k-1} \overset{a.s.}{=} \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t \varepsilon'_t + \mathrm{o}(T^{-\xi}),$$

If in addition Assumption (2.2) is satisfied then for all $\zeta < \min(\xi, 1/2)$ it holds

$$\hat{\Omega}_{k-1} \overset{a.s.}{=} \Omega + \mathrm{o}(T^{-\zeta}).$$

3.3. Asymptotic analysis of $\hat{\Omega}_{k-1} - \hat{\Omega}_k$

The analysis of the term $\hat{\Omega}_{k-1} - \hat{\Omega}_k$ is specific to the order selection problem. For the sake of finding the asymptotic distribution of the likelihood ratio test statistic the aim is to express $\hat{\Omega}_{k-1} - \hat{\Omega}_k$ in terms of a stationary process Y_t as

$$T(\hat{\Omega}_{k-1} - \hat{\Omega}_k) = Q(X_{t-k}|\mathbf{X}_{t-1}, D_t) = Q(Y_{t-1}) + \mathrm{o}_{\mathsf{P}}(1), \tag{3.3}$$

which in turn can be proved to be asymptotically χ^2 by a Central Limit Theorem. The result (3.3) reduces trivially to an equality with $Y_{t-1} = \varepsilon_{t-1}$ when testing $A_1 = 0$, so only the case $k > 1$ will need consideration in the remainder of this subsection. On the way to prove the above result some related expressions holding under weaker assumptions emerge which can be used for proving the consistency results for the estimator of the lag length, $\hat{k}$.

In the following $\hat{\Omega}_{k-1} - \hat{\Omega}_k$ is first decomposed into seven terms. It is then shown that the three leading term can be written as $Q(Y_{t-1})$ as in (3.3) and that the remaining four terms are asymptotically vanishing.

3.3.1. Decomposition of $\hat{\Omega}_{k-1} - \hat{\Omega}_k$

The first decomposition is a purely algebraic result based on the formula for partitioned inversion.

Lemma 3.2. *Suppose* $A_k = 0$. *Then it holds*

$$Q\left(X_{t-k}|\mathbf{X}_{t-1}, D_t\right) = Q\left(\mathbf{X}_{t-2}|D_t\right) - Q\left(\mathbf{X}_{t-1}|D_t\right) + Q\left(\varepsilon_{t-1}|\mathbf{X}_{t-2}, D_t\right).$$

Proof of Lemma 3.2. By the formula for partitioned inversion it holds

$$Q\left(\begin{array}{c}\mathbf{X}_{t-1}\\ X_{t-k}\end{array}\middle| D_t\right) = Q\left(X_{t-k}|\mathbf{X}_{t-1}, D_t\right) + Q\left(\mathbf{X}_{t-1}|D_t\right), \tag{3.4}$$

of which $T(\hat{\Omega}_{k-1} - \hat{\Omega}_k) = Q\left(X_{t-k}|\mathbf{X}_{t-1}, D_t\right)$ is the first term on the left. Noting that $(\mathbf{X}'_{t-1}, X'_{t-k})' = (X'_{t-1}, \mathbf{X}'_{t-2})'$ a repeated use of the formula for partitioned inversion shows

$$Q\left(\begin{array}{c}\mathbf{X}_{t-1}\\ X_{t-k}\end{array}\middle| D_t\right) = Q\left(\begin{array}{c}X_{t-1}\\ \mathbf{X}_{t-2}\end{array}\middle| D_t\right) = Q\left(X_{t-1}|\mathbf{X}_{t-2}, D_t\right) + Q\left(\mathbf{X}_{t-2}|D_t\right). \tag{3.5}$$

Due to the model equation (1.1) with $A_k = 0$ and the property $D_t = \mathbf{D}D_{t-1}$ it follows $(X_{t-1}|\mathbf{X}_{t-2}, D_t) = (\varepsilon_{t-1}|\mathbf{X}_{t-2}, D_t)$. The desired expression then arise by rearranging the above expressions. □

Asymptotic arguments are now needed. These arguments rely on Nielsen [17] which in turn represents a generalisation of the arguments of Lai and Wei [15]. The second step is therefore an asymptotic decomposition of the first two terms in Lemma 3.2 using that the processes U_t, V_t, W_t are asymptotically uncorrelated.

Lemma 3.3. *Suppose* $A_k = 0$ *and that the Assumptions* (2.1), (2.3), (2.5) *are satisfied with* $\lambda > 2$. *Then, for* $j = 1, 2$,

$$Q\left(\mathbf{X}_{t-j}|D_t\right) \overset{a.s.}{=} Q\left(U_{t-j}|D_t\right) + Q\left(V_{t-j}|D_t\right) + Q\left(W_{t-j}|D_t\right) + \mathrm{o}\left(1\right). \tag{3.6}$$

Proof of Lemma 3.3. Since $M\mathbf{X}_t = (U_t, V_t, W_t)$, see (2.7), it suffices to argue that the processes U_t, V_t and W_t are asymptotically uncorrelated so that the off-diagonal elements of $\sum_{t=1}^{T}(\mathbf{X}_{t-j}|D_t)(\mathbf{X}_{t-j}|D_t)'$ can be ignored in the asymptotic argument. This follows from Nielsen ([17], Theorem 6.4, 9.1, 9.2, 9.4), see also the summary in Table 2 of that paper. □

3.3.2. Eliminating explosive terms and regressors in stationary terms

In combination Lemmas 3.2, 3.3 show that

$$\begin{aligned} T(\hat{\Omega}_{k-1} - \hat{\Omega}_k) \overset{a.s.}{=}\ & Q\left(\varepsilon_{t-1}|\mathbf{X}_{t-2}, D_t\right) + Q\left(U_{t-2}|D_t\right) - Q\left(U_{t-1}|D_t\right) \\ & + Q\left(V_{t-2}|D_t\right) - Q\left(V_{t-1}|D_t\right) + Q\left(W_{t-2}|D_t\right) - Q\left(W_{t-1}|D_t\right) + \mathrm{o}\left(1\right). \end{aligned}$$

Under mild conditions this can be reduced further so as to eliminate the terms involving the explosive component W_t as well as the regressors in the terms involving the stationary component U_t.

Lemma 3.4. *Suppose $A_k = 0$ and that the Assumptions (2.1), (2.3), (2.5) are satisfied, with $\lambda > 2$. Then,*

$$T(\hat{\Omega}_{k-1} - \hat{\Omega}_k) \overset{a.s.}{=} Q(\varepsilon_{t-1}) + Q(\tilde{U}_{t-2}) - Q(\tilde{U}_{t-1}) + R_\varepsilon + R_V + \mathrm{o}(1), \tag{3.7}$$

where

$$R_\varepsilon = Q(\varepsilon_{t-1}|\mathbf{X}_{t-2}, D_t) - Q(\varepsilon_{t-1}), \quad R_V = Q(V_{t-2}|D_t) - Q(V_{t-1}|D_t). \tag{3.8}$$

Proof of Lemma 3.4. It suffices to prove, for $j = 1, 2$,

$$Q(U_{t-j}|D_t) \overset{a.s.}{=} Q(\tilde{U}_{t-j}) + \mathrm{o}(1), \tag{3.9}$$

$$Q(W_{t-2}|D_t) - Q(W_{t-1}|D_t) \overset{a.s.}{=} \mathrm{o}(1). \tag{3.10}$$

First, consider (3.9). Because of (2.8) then $(U_{t-j}|D_t) = (\tilde{U}_{t-j}|D_t)$. According to Nielsen ([17], Theorem 6.4) it holds for any $\eta > 0$ that

$$\left(\sum_{t=1}^{T} D_t D_t'\right)^{-1/2} \sum_{t=1}^{T} D_t \tilde{U}_{t-j}' \left(\sum_{t=1}^{T} \tilde{U}_{t-j}\tilde{U}_{t-j}'\right)^{-1/2} \overset{a.s.}{=} \mathrm{o}(T^{\eta-1/2}),$$

while Theorem 6.2 of the above paper shows $T^{-1}\sum_{t=1}^{T} \tilde{U}_{t-j}\tilde{U}_{t-j}'$ has positive definite limit points. This implies

$$\sum_{t=1}^{T} (\tilde{U}_{t-j}|D_t)(\tilde{U}_{t-j}|D_t)' \overset{a.s.}{=} \sum_{t=1}^{T} \tilde{U}_{t-j}\tilde{U}_{t-j}' \left\{1 + \mathrm{o}\left(T^{2\eta-1}\right)\right\}.$$

Theorem 2.4 of the above paper shows $\sum_{t=1}^{T} \varepsilon_t D_t' (\sum_{t=1}^{T} D_t D_t')^{-1/2} = \mathrm{o}(T^\eta)$ implying

$$\sum_{t=1}^{T} \varepsilon_t (\tilde{U}_{t-j}|D_t)' \overset{a.s.}{=} \sum_{t=1}^{T} \varepsilon_t \tilde{U}_{t-j}' + \mathrm{o}(T^{2\eta}).$$

That theorem also shows $\sum_{t=1}^{T} \varepsilon_t \tilde{U}_{t-j}' (\sum_{t=1}^{T} \tilde{U}_{t-j}\tilde{U}_{t-j}')^{-1/2} = \mathrm{o}(T^\eta)$. In combination these results show the desired result.

Secondly, consider (3.10). Note first that $W_{t-1} = \mathbf{W}W_{t-2} + \mu_W D_{t-1} + e_{W,t-1}$ by (2.7) while $D_{t-1} = \mathbf{D}^{-1}D_t$, implying $(W_{t-1}|D_t) = (\mathbf{W}W_{t-2} + e_{W,t-1}|D_t)$. This gives rise to the expansions

$$\sum_{t=1}^{T} (W_{t-1}|D_t)^{\otimes 2} = \sum_{t=1}^{T} (\mathbf{W}W_{t-2}|D_t)^{\otimes 2} (1 + f_T),$$

$$\sum_{t=1}^{T} (W_{t-1}|D_t)\, \varepsilon_t = \sum_{t=1}^{T} (\mathbf{W}W_{t-2}|D_t)\, \varepsilon_t + c_T,$$

where $f_T = \mathrm{O}(d_T^{-1/2} a_T) + d_T^{-1} b_T$ and

$$a_T = d_T^{-1/2} \sum_{t=1}^{T} (\mathbf{W}W_{t-2}|D_t)\, e_{W,t-1}, \qquad b_T = \sum_{t=1}^{T} (e_{W,t-1}|D_t)^{\otimes^2},$$

$$c_T = \sum_{t=1}^{T} (e_{W,t-1}|D_t)\, \varepsilon_t, \qquad d_T = \sum_{t=1}^{T} (\mathbf{W}W_{t-2}|D_t)^{\otimes 2}.$$

Using Nielsen ([17], Theorems 2.4, 6.2, 6.4) it is seen that

$$b_T \overset{a.s.}{=} \mathrm{O}(T), \quad c_T \overset{a.s.}{=} \mathrm{o}(T^{1/2+\eta}).$$

It follows from Nielsen ([17], Theorems 2.4, 9.1 and Corollary 7.2) that

$$Q\left(W_{t-j}|D_t\right) \overset{a.s.}{=} \mathrm{o}\left(T\right), \quad a_T \overset{a.s.}{=} \mathrm{o}(T^{1/2}), \quad d_T^{-1} \overset{a.s.}{=} \mathrm{o}\left(\rho^{-T}\right),$$

for some $\rho > 0$. This implies that f_T is exponentially decreasing. The desired result follows by expanding $Q(W_{t-1}|D_t)$ in terms of $Q(W_{t-2}|D_t)$ as

$$\left[Q\left(W_{t-2}|D_t\right) + d_T^{-1/2} c_T \mathrm{O}\left\{Q\left(W_{t-2}|D_t\right)\right\}^{1/2} + c_T' d_T^{-1} c_T\right]\left(1 + f_T\right),$$

and using the established orders of magnitude. □

3.3.3. Eliminating unit root terms and regressors in innovation terms

The terms R_V and R_ε defined in (3.8) are now shown to vanish asymptotically. At first, consider R_V defined in (3.8), which consists of the terms involving the unit root components V_t. Several results are given, of which the strongest result for R_V can only be established for certain values of the parameters.

Lemma 3.5. *Suppose* $A_k = 0$ *and that the Assumptions* (2.1), (2.3), (2.5) *are satisfied with* $\lambda > 2$. *Then*

(i) $R_V \overset{a.s.}{=} \mathrm{O}(\log T)$,
(ii) $R_V = \mathrm{o_P}(1)$ *if also Assumption* (2.2) *holds,*
(iii) $R_V \overset{a.s.}{=} \mathrm{O}\{(\log\log T)^{1/2}(\log T)^{1/2}\}$ *if* (*A*) $\mathbf{D}$ *and* $\mathbf{V}$ *have no common eigenvalues,*
(iv) $R_V \overset{a.s.}{=} \mathrm{o}(1)$ *if* (*B*) $\dim \mathbf{D} = 0$ *and* $\mathbf{V} = 1$,
(v) $R_V = 0$ *if* (*C*) $\dim \mathbf{V} = 0$.

Proof of Lemma 3.5. (i) This follows since $Q(V_{t-j}|D_t) \overset{a.s.}{=} \mathrm{O}(\log T)$ according to Nielsen ([17], Theorem 2.4).

(ii) The type of argument for (3.10) in the proof of Lemma 3.4 can be used. Replacing W with V throughout, the asymptotic properties of a_T, b_T, c_T, d_T have to be explored. For b_T, c_T the argument is the same so, for all $\eta > 0$,

$$b_T \overset{a.s.}{=} \mathrm{O}(T), \qquad c_T \overset{a.s.}{=} \mathrm{o}(T^{1/2+\eta}),$$

whereas using Nielsen ([17], Theorems 2.4) for a_T and the techniques of Chan and Wei [5] for d_T shows, for all $\eta > 0$,

$$a_T \overset{a.s.}{=} \mathrm{o}\left(T^{\eta}\right), \qquad d_T^{-1} = \mathrm{o_P}(T^{-1-4\eta}),$$

so $f_T = \mathrm{o_P}(T^{-4\eta})$. Since $Q(V_{t-j}|D_t) \overset{a.s.}{=} \mathrm{O}(\log T)$ as established in (i) the desired result follows by expanding $Q(V_{t-1}|D_t)$ in terms of $Q(V_{t-2}|D_t)$.

(iii) Define the vector $S_{t-1} = (V_{t-1}', D_t')'$. By partitioned inversion it holds

$$Q\left(S_{t-1}\right) = Q\left(V_{t-1}|D_t\right) + Q\left(D_t\right).$$

By an invariance argument D_t can be replaced by D_{t-j} and thus it follows

$$R_V = Q\left(V_{t-2}|D_t\right) - Q\left(V_{t-1}|D_t\right) = Q\left(S_{t-2}\right) - Q\left(S_{t-1}\right).$$

Due to (2.4) and (2.7) the process S_{t-1} satisfies $S_t = \mathbf{S}S_{t-1} + e_{S,t}$ for a matrix $\mathbf{S}$ with eigenvalues of length one and $e_{S\,t} = (e'_{V,t}, 0')'$. It then follows that

$$\sum_{t=1}^{T} \varepsilon_t S'_{t-1} = \sum_{t=1}^{T} \varepsilon_t \left(S'_{t-2}\mathbf{S}' + e'_{S,t-1} \right).$$

Inserting this expression into $Q(S_{t-1})$ shows

$$Q\left(S_{t-1}\right) = \sum_{t=1}^{T} \varepsilon_t S'_{t-1} \left(\sum_{t=1}^{T} S_{t-1}^{\otimes 2} \right)^{-1} \sum_{t=1}^{T} S_{t-1}\varepsilon'_t = Q_A + Q_B + Q_C + Q'_C,$$

where

$$Q_A = Q_1 Q_2 Q'_1, \quad Q_B = Q_4 Q_3 Q'_3 Q'_4, \quad Q_C = Q_1 Q_2^{1/2} Q'_3 Q'_4,$$

are defined in terms of the statistics

$$Q_1 = \sum_{t=1}^{T} \varepsilon_t e'_{S,t-1}, \quad Q_2 = \left(\sum_{t=1}^{T} S_{t-1}^{\otimes 2} \right)^{-1},$$
$$Q_3 = \left(\sum_{t=1}^{T} S_{t-2}^{\otimes 2} \right)^{1/2} \mathbf{S} \left(\sum_{t=1}^{T} S_{t-1}^{\otimes 2} \right)^{-1/2}, \quad Q_4 = \sum_{t=1}^{T} \varepsilon_t S'_{t-2} \left(\sum_{t=1}^{T} S_{t-2}^{\otimes 2} \right)^{-1/2}.$$

The orders of magnitude of these follow from a series of results in Nielsen [17]. Theorem 6.1 and Lemma 6.3 imply $Q_1 \overset{a.s.}{=} \mathrm{O}\{(T \log\log T)^{1/2}\}$. Theorem 8.3 shows $Q_2 \overset{a.s.}{=} \mathrm{O}(T^{-1})$ when $\mathbf{D}$ and $\mathbf{V}$ have no common eigenvalues. Lemma 8.7(*ii*) shows $Q_3^{\otimes 2} - I \overset{a.s.}{=} \mathrm{O}\{T^{-1/2}(\log T)^{1/2}\}$. Theorem 2.4 shows $Q_4 \overset{a.s.}{=} \mathrm{O}\{(\log T)^{1/2}\}$. Noting that $Q(S_{t-2}) = Q_4 Q'_4$ this in turn implies

$$Q_A = \mathrm{O}(\log\log T), \qquad Q_B = Q(S_{t-2}) + \mathrm{O}\{T^{-1/2}(\log T)^{3/2}\},$$
$$Q_C = \mathrm{O}\{(\log\log T)^{1/2}(\log T)^{1/2}\},$$

and the desired result follows.

(iv) Donsker and Varadhan's [6] Law of the Iterated Logarithm for the integrated squared Brownian motion states

$$\liminf_{T\to\infty} \frac{\log\log T}{T^2} \int_0^T B_u^2 du \overset{a.s.}{=} \frac{1}{4}.$$

Now use *either* the argument in (ii) with $d_T^{-1} \overset{a.s.}{=} \mathrm{O}(T^{-2}\log\log T)$ *or* the argument in (iii) with $Q_2 \overset{a.s.}{=} \mathrm{O}(T^{-2}\log\log T)$ so Q_A, Q_B, Q_C are all o(1).

(v) This follows by construction. □

Now, consider R_ε defined in (3.8). By showing that this vanishes it follows that the regressors can be excluded asymptotically in the term involving the lagged innovations ε_{t-1}. A fourth order moment condition is now needed in Assumption (2.1).

Lemma 3.6. *Suppose* $A_k = 0$ *and that the Assumptions* (2.1), (2.3), (2.5) *are satisfied, now with* $\lambda > 4$. *Then*

$$R_\varepsilon = Q\left(\varepsilon_{t-1} | \mathbf{X}_{t-2}, D_t\right) - Q\left(\varepsilon_{t-1}\right) \overset{a.s.}{=} \mathrm{o}\left(1\right).$$

Proof of Lemma 3.6. Define the vector $S_t = (\mathbf{X}'_{t-2}, D'_t)'$. According to Nielsen ([17], Theorem 2.4) it holds that, for any $\eta > 0$, the terms

$$\text{(3.11)} \qquad \left(\sum_{t=1}^{T} S_t S'_t\right)^{-1/2} \sum_{t=1}^{T} S_t \varepsilon'_t, \qquad \left(\sum_{t=1}^{T} S_t S'_t\right)^{-1/2} \sum_{t=1}^{T} S_t \varepsilon'_{t-1}$$

are $\mathrm{o}(T^{1/4-\eta})$ when indeed $\lambda > 4$. It then holds that

$$\sum_{t=1}^{T} \varepsilon_t \varepsilon'_{t-1} - \sum_{t=1}^{T} \varepsilon_t S'_t \left(\sum_{t=1}^{T} S_t S'_t\right)^{-1} \sum_{t=1}^{T} S_t \varepsilon'_{t-1} \overset{a.s.}{=} \sum_{t=1}^{T} \varepsilon_t \varepsilon'_{t-1} + \mathrm{o}(T^{1/2-\eta}),$$

$$\sum_{t=1}^{T} \varepsilon_{t-1} \varepsilon'_{t-1} - \sum_{t=1}^{T} \varepsilon_{t-1} S'_t \left(\sum_{t=1}^{T} S_t S'_t\right)^{-1} \sum_{t=1}^{T} S_t \varepsilon'_{t-1} \overset{a.s.}{=} \sum_{t=1}^{T} \varepsilon_{t-1} \varepsilon'_{t-1} + \mathrm{o}(T^{1-\eta}),$$

where the requirement $\lambda > 4$ is only needed in the first case. Theorems 2.5, 6.1 of the above paper show $T^{-1} \sum_{t=1}^{T} \varepsilon_{t-1} \varepsilon'_{t-1}$ has positive definite limit points while $\sum_{t=1}^{T} \varepsilon_t \varepsilon'_{t-1} (\sum_{t=1}^{T} \varepsilon_{t-1} \varepsilon'_{t-1})^{-1/2} = \mathrm{o}(T^{\eta})$. Combine these results. □

Remark 3.7. In Lemma 3.6 a fourth moment condition comes in through the requirement that $\lambda > 4$ in Assumption (2.1). This can be relaxed to $\lambda > 2$ under one of two alternative assumptions.

(I,a) If $\dim \mathbf{W} = 0$ then the terms in (3.11) are $\mathrm{o}(T^{\eta})$, see Nielsen ([17], Theorem 2.4), and the main result holds.

(I,b) If $\dim \mathbf{W} > 0$ but the innovations ε_t are independently, identically distributed then terms of the type $(\sum_{t=1}^{T} W_{t-1} W'_{t-1})^{-1/2} \sum_{t=1}^{T} W_{t-1} \varepsilon'_t$ converge in distribution, see Anderson [1] and the result of the Theorem holds, albeit only in probability.

3.3.4. The leading term of $\hat{\Omega}_{k-1} - \hat{\Omega}_k$

First the order of magnitude the leading term in (3.7) is established in an almost sure sense. This can be done under weak moment conditions. Subsequently the distribution of the leading term is investigated.

Lemma 3.8. *Suppose* $A_k = 0$ *and that the Assumptions* (2.1), (2.3) *are satisfied with* $\lambda > 2$. *Define* $E_T = T^{-1} \sum_{t=1}^{T} \varepsilon_t \varepsilon'_t$. *Then*
$\limsup_{T \to \infty} (2 \log \log T)^{-1} \mathrm{tr}[\{Q(\varepsilon_{t-1}) + Q(\tilde{U}_{t-2}) - Q(\tilde{U}_{t-1})\} E_T^{-1}] \overset{a.s.}{=} \mathrm{O}(1)$.

Proof of Lemma 3.8. This follows by noting that the sequence $\hat{\Omega}_{k-1}^{-1}$ is relatively compact with positive definite limiting points due to Lemma 3.1 and Lai and Wei ([15], Theorem 2) and otherwise following the argument in the proof of Pötscher ([21], Theorem 3.4). □

When it comes to analysing the distribution of the leading term in (3.7) it is convenient to show that it can be written as a single quadratic form $Q(Y_{t-1})$ for some process Y_{t-1}. This argument requires two steps, of which the first is concerned with the convergence properties of $T^{-1} \sum_{t=1}^{T} \tilde{U}_{t-1} \tilde{U}'_{t-1}$. As the argument involves a variance matrix, the Assumption (2.2) is now called upon.

Lemma 3.9. *Suppose $A_k = 0$ and that the Assumptions (2.1), (2.2) are satisfied with $\lambda > 2$. Let M_U be the matrix defined by $e_{U,t} = M_U \varepsilon_t$ in (2.7) and define*

$$F = \sum_{t=0}^{\infty} \mathbf{U}^t M_U \Omega M_U' (\mathbf{U}^t)'.$$

Then for all $\zeta < \min(1 - 2/\lambda, 1/2)$ it holds

$$\frac{1}{T} \sum_{t=1}^{T} \tilde{U}_t \tilde{U}_t' \overset{a.s.}{=} F + \mathrm{o}\left(T^{-\zeta}\right).$$

Proof of Lemma 3.9. Following the proof of Lai and Wei ([15], Theorem 2), the equation (2.8) shows

$$\sum_{t=1}^{T} \tilde{U}_t \tilde{U}_t' \overset{a.s.}{=} \mathbf{U} \left(\sum_{t=1}^{T} \tilde{U}_t \tilde{U}_t' - \tilde{U}_T \tilde{U}_T' + \tilde{U}_0 \tilde{U}_0' \right) \mathbf{U}' + M_U \sum_{t=1}^{T} \varepsilon_t \varepsilon_t' M_U' + \mathrm{O}\left(\sum_{t=1}^{T} \tilde{U}_{t-1} \varepsilon_t' \right).$$

Due to Nielsen ([17], Theorems 2.4, 5.1, Example 6.5) both $\sum_{t=1}^{T} \tilde{U}_{t-1} \varepsilon_t'$ and $\tilde{U}_T \tilde{U}_T'$ are $\mathrm{o}(T^{1-\zeta})$. Note that Assumption (2.5) is not needed as $\tilde{U}_t$ does not involve deterministic terms. Denoting $F_T = T^{-1} \sum_{t=1}^{T} \tilde{U}_t \tilde{U}_t'$ it follows from Lemma 3.1 that

$$F_T - \mathbf{U} F_T \mathbf{U}' \overset{a.s.}{=} M_U \Omega M_U' + \mathrm{o}(T^{-\zeta}).$$

This equation has a unique solution $F_T = \sum_{t=0}^{\infty} \mathbf{U}^t \{ M_U E M_U' + \mathrm{o}(T^{-\zeta}) \} (\mathbf{U}^t)'$, see Anderson and Moore ([2], p. 336), which in turn equals $F + \mathrm{o}(T^{-\zeta})$ since the maximal eigenvalue of $\mathbf{U}\mathbf{U}'$ is less than one. □

The leading term in (3.7) is now written as a single quadratic form $Q(Y_{t-1})$.

Lemma 3.10. *Suppose $A_k = 0$ and that the Assumptions (2.1), (2.2) are satisfied with $\lambda > 2$. Then there exists an $\{(p + \dim U) \times p\}$-matrix C with full column rank so*

$$Q\left(\varepsilon_{t-1}\right) + Q(\tilde{U}_{t-2}) - Q(\tilde{U}_{t-1}) \overset{a.s.}{=} Q\left(Y_{t-1}\right) + \mathrm{o}\left(1\right),$$

where Y_t is the process $C'(\varepsilon_t', U_{t-1}')'$.

Proof of Lemma 3.10. The idea is to exploit that the asymptotic covariance for $Z_{t-1} = (\tilde{U}_{t-2}', \varepsilon_{t-1}')'$ is diagonal with elements F, Ω. By the above Lemmas 3.1, 3.9 then, for some $\eta > 0$,

$$\begin{aligned} & Q\left(\varepsilon_{t-1}\right) + Q(\tilde{U}_{t-2}) \\ & \quad = \sum_{t=1}^{T} \varepsilon_t \begin{pmatrix} \tilde{U}_{t-2} \\ \varepsilon_{t-1} \end{pmatrix}' \left\{ \sum_{t=1}^{T} \begin{pmatrix} \tilde{U}_{t-2} \tilde{U}_{t-2}' & 0 \\ 0 & \varepsilon_{t-1} \varepsilon_{t-1}' \end{pmatrix} \right\}^{-1} \sum_{t=1}^{T} \begin{pmatrix} \tilde{U}_{t-2} \\ \varepsilon_{t-1} \end{pmatrix} \varepsilon_t' \\ & \quad \overset{a.s.}{=} \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t \begin{pmatrix} \tilde{U}_{t-2} \\ \varepsilon_{t-1} \end{pmatrix}' \begin{pmatrix} F & 0 \\ 0 & \Omega \end{pmatrix}^{-1} \sum_{t=1}^{T} \begin{pmatrix} \tilde{U}_{t-2} \\ \varepsilon_{t-1} \end{pmatrix} \varepsilon_t' \{1 + \mathrm{o}(T^{-\eta})\} \end{aligned} \tag{3.12}$$

As discussed in Section 2 then $\tilde{U}_{t-1} = \mathbf{U} \tilde{U}_{t-2} + M_U \varepsilon_{t-1}$ for some matrix M_U with full column rank. In particular $\tilde{U}_{t-1} = C_\perp' (\tilde{U}_{t-2}', \varepsilon_{t-1}')'$ where the $\{(p + \dim U) \times$

$\dim U\}$-matrix $C_\perp = (\mathbf{U}, M_U)'$ has full column rank. Therefore a $\{(p+\dim U)\times p\}$-matrix C can be chosen with full column rank so the matrix $(C, C_\perp)$ is regular and

$$C' \begin{pmatrix} F & 0 \\ 0 & \Omega \end{pmatrix} C_\perp = 0.$$

The sequences $T^{-1}\sum_{t=1}^{T} \tilde{U}_{t-1}\tilde{U}'_{t-1}$ and $T^{-1}\sum_{t=1}^{T} \tilde{U}_{t-2}\tilde{U}'_{t-2}$ will have the same limit, F, while $T^{-1}\sum_{t=1}^{T} Y_{t-1}Y'_{t-1}$ will converge to a positive definite matrix G. It then holds

$$\begin{pmatrix} C'_\perp \\ C' \end{pmatrix} \begin{pmatrix} F & 0 \\ 0 & \Omega \end{pmatrix} (C_\perp, C) = \begin{pmatrix} F & 0 \\ 0 & G \end{pmatrix}.$$

Pre- and post-multiplying the middle matrix in (3.12) with $(C_\perp, C)(C_\perp, C)^{-1}$ and its transpose then implies

$$Q\left(\varepsilon_{t-1}\right) + Q(\tilde{U}_{t-2}) \overset{a.s.}{=} \left\{Q(\tilde{U}_{t-1}) + Q\left(Y_{t-1}\right)\right\}\{1 + \mathrm{o}(T^{-\eta})\}.$$

Theorem 2.4 of Nielsen (2005) implies $Q(\tilde{U}_{t-1})$ and $Q(Y_{t-1})$ are $\mathrm{o}(T^{\eta})$, which gives the desired result. □

The asymptotic distribution of the leading term $Q(Y_{t-1})$ now follows.

Lemma 3.11. *Suppose* $A_k = 0$ *and that the Assumptions* (2.1), (2.2), (2.3) *are satisfied with* $\lambda > 4$. *Then*

(i) $1 \leq \limsup_{T\to\infty}(2\log\log T)^{-1}\mathrm{tr}\{Q(Y_{t-1})\Omega^{-1}\} \leq p^2$ *a.s.*

(ii) $\mathrm{tr}\{Q(Y_{t-1})\Omega^{-1}\} \xrightarrow{\mathsf{D}} \chi^2(p^2)$.

Proof of Lemma 3.11. (i) This follows from the Law of Iterated Logarithms by Heyde and Scott ([12], Corollary 2) and Hannan ([9], p. 1076-1077). See Quinn [22] for details.

(ii) This follows from Brown and Eagleson's [4] Central Limit Theorem. This requires existence of second moments of $\varepsilon_t Y_{t-1}$. □

Remark 3.12. The proof of Lemma 3.11 actually only requires the existence of fourth moments, which is slightly weaker than the stated condition of $\lambda > 4$ in Assumption (2.1). In Lemma 3.11(ii) this can be relaxed to a second moment condition if for instance:
(II) the innovations ε_t are independent.

3.4. Proofs of results for likelihood ratio test statistics

Proof of Theorem 2.1. Consider the formula (3.1). The term $\hat{\Omega}_{k-1}$ was dealt with in Lemma 3.1. As for the term $T(\hat{\Omega}_{k-1} - \hat{\Omega}_k)$ consider two cases.

When $k = 1$ then $T(\hat{\Omega}_{k-1} - \hat{\Omega}_k) = Q(\varepsilon_{t-1})$.

When $k > 1$ apply the expansion in Lemma 3.4. The term R_V vanishes due to Lemma 3.5(ii) when Assumption (2.2) is satisfied. The term R_ε vanishes due to Lemma 3.6 when $\lambda > 4$ in Assumption (2.1). Due to Lemma 3.10 the leading term is now $Q(Y_{t-1})$, provided Assumption (2.2) holds.

For any k the desired χ^2-distribution now arises from Lemma 3.11(ii) provided Assumptions (2.2), (2.1) are satisfied with $\lambda > 4$. □

Proof of Theorem 2.3. Note first that $T(\hat{\Omega}_{k-1} - \hat{\Omega}_{k+m-1})$ can be written as $Q(\tilde{X}_{t-k}^{t-k-m+1}|\mathbf{X}_{t-1}, D_t)$ where $\tilde{X}_{t-a}^{t-b} = (X'_{t-a}, \ldots, X'_{t-b})'$. Consider now the proof of the decomposition in Lemma 3.2. Using first (3.4) and then (3.5) repeatedly it is seen that

$$\begin{aligned} T\left(\hat{\Omega}_{k-1} - \hat{\Omega}_{k+m-1}\right) &= Q\left(\left.\begin{array}{c}\mathbf{X}_{t-1}\\ \tilde{X}_{t-k}^{t-k-m+1}\end{array}\right| D_t\right) - Q\left(\mathbf{X}_{t-1}|\, D_t\right) \\ &= \sum_{j=1}^{m} Q\left(\varepsilon_{t-j}|\, X_{t-j-1}^{t-m}, \mathbf{X}_{t-m-1}, D_t\right) \\ &\quad + Q\left(\mathbf{X}_{t-m-1}|\, D_t\right) - Q\left(\mathbf{X}_{t-1}|\, D_t\right). \end{aligned}$$

As in the proof of Theorem 2.1 the Lemmas 3.4, 3.5(ii), 3.6 show that the leading terms reduce to

$$T\left(\hat{\Omega}_{k-1} - \hat{\Omega}_{k+m-1}\right) = \sum_{j=1}^{m} Q\left(\varepsilon_{t-j}\right) + Q\left(\tilde{U}_{t-m-1}\right) - Q\left(\tilde{U}_{t-1}\right) + \mathrm{o_P}\left(1\right),$$

when $k_0 < k$. A slight generalisation of Lemma 3.10 is needed, using that the asymptotic covariance for $Z_{t-1} = (\tilde{U}'_{t-m-1}, \varepsilon'_{t-1}, \ldots, \varepsilon'_{t-m})'$ is diagonal with elements $F, \Omega, \ldots, \Omega$. A $\{(mp + \dim U) \times mp\}$-matrix C can then be found giving rise to a process $Y_{t-1} = C' Z_{t-1}$. The argument is completed using a Central Limit Theorem as in the proof of Lemma 3.11(ii). □

3.5. Proofs of results for information criteria

Proof of Theorem 2.4. Consider $j < k_0$. The condition $f(T) = \mathrm{o}(T)$ implies

$$\Phi_j - \Phi_{k_0} = \log\det\{I + (\hat{\Omega}_j - \hat{\Omega}_{k_0})\hat{\Omega}_{k_0}^{-1}\} + \mathrm{o}\left(1\right).$$

Lemma 3.1 shows that $\hat{\Omega}_{k_0} \overset{a.s.}{\to} \Omega$, so it suffices that $\liminf_{T\to\infty} \lambda_{\max}(\hat{\Omega}_j - \hat{\Omega}_{k_0})$ is positive. Defining $\mathbf{Y}_t = (X'_{t-1}, \ldots, X'_{t-j+1})'$ and $\mathbf{Z}_t = (X'_{t-j}, \ldots, X'_{t-k_0})'$ it holds

$$\hat{\Omega}_j - \hat{\Omega}_{k_0} = \left[T^{-1/2} \sum_{t=1}^{T} X_t \left(\mathbf{Z}_{t-1}|\mathbf{Y}_{t-1}, D_t\right)' \left\{\sum_{t=1}^{T} \left(\mathbf{Z}_{t-1}|\mathbf{Y}_{t-1}, D_t\right)^{\otimes 2}\right\}^{-1/2}\right]^{\otimes 2}.$$

Define $\mathbf{A_y} = A_1, \ldots, A_j$ and $\mathbf{A_z} = A_{j+1}, \ldots, A_{k_0}$ noting that $A_{k_0} \neq 0$. Then it holds $X_t = \mathbf{A_y}\mathbf{Y}_t + \mathbf{A_z}\mathbf{Z}_t + \mu D_t + \varepsilon_t$. Therefore $\hat{\Omega}_j - \hat{\Omega}_{k_0}$ equals

$$\begin{aligned} &T^{-1/2} \sum_{t=1}^{T} \varepsilon_t \left(\mathbf{Z}_{t-1}|\mathbf{Y}_{t-1}, D_t\right)' \left\{\sum_{t=1}^{T} \left(\mathbf{Z}_{t-1}|\mathbf{Y}_{t-1}, D_t\right)^{\otimes 2}\right\}^{-1/2} \\ &+\mathbf{A_z} \left\{T^{-1} \sum_{t=1}^{T} \left(\mathbf{Z}_{t-1}|\mathbf{Y}_{t-1}, D_t\right)^{\otimes 2}\right\}^{1/2}. \end{aligned}$$

The first term is of order o(1) *a.s.* by Nielsen ([17], Theorem 2.4). As for the second term it holds that $\liminf_{T\to\infty} \lambda_{\min}\{T^{-1} \sum_{t=1}^{T} (\mathbf{X}_{t-1}|D_t)^{\otimes 2}\} > 0$ *a.s.* according to Nielsen ([17], Corollary 9.5). As a consequence the limit points of $T^{-1} \sum_{t=1}^{T} (\mathbf{Z}_{t-1}|\, \mathbf{Y}_{t-1}, D_t)^{\otimes 2}$ are positive definite. Since $\mathbf{A}_z \neq 0$ then $\liminf_{T\to\infty} \lambda_{\min}(\hat{\Omega}_j - \hat{\Omega}_{k_0}) > 0$ and therefore $\liminf_{T\to\infty} \hat{k} \geq k_0$ *a.s.* □

Proof of Theorem 2.5. Consider now $k_0 < j \leq K$. It then holds

$$\begin{aligned}\Phi_{j+1} - \Phi_j &= \log\det(\hat{\Omega}_{j+1}\hat{\Omega}_j^{-1}) + T^{-1}f(T) \\ &= \log\det\{I_p - (\hat{\Omega}_j - \hat{\Omega}_{j+1})\hat{\Omega}_j^{-1}\} + T^{-1}f(T).\end{aligned}$$

A Taylor expansion shows

$$\Phi_{j+1} - \Phi_j \overset{a.s.}{=} -\mathrm{tr}\{(\hat{\Omega}_j - \hat{\Omega}_{j+1})\hat{\Omega}_j^{-1}\} + T^{-1}f(T) + \mathrm{o}[\{(\hat{\Omega}_j - \hat{\Omega}_{j+1})\hat{\Omega}_j^{-1}\}^2].$$

Lemma 3.1 shows that $\hat{\Omega}_j$ is consistent, while Lemma 3.4 gives the expansion

$$T(\hat{\Omega}_{j-1} - \hat{\Omega}_j) \overset{a.s.}{=} Q(\varepsilon_{t-1}) + Q(\tilde{U}_{t-2}) - Q(\tilde{U}_{t-1}) + R_\varepsilon + R_V + \mathrm{o}(1).$$

To complete the proof it has to be shown that $\Phi_{j+1} - \Phi_j$ has a positive limiting value. This holds if $T(\hat{\Omega}_{j-1} - \hat{\Omega}_j) = \mathrm{o}\{g(T)\}$ for some function $g(T)$ so $f(T)/g(T) \to \infty$.

(i) The term R_V vanishes due to Lemma 3.5(ii) when Assumption (2.2) is satisfied. The term R_ε vanishes due to Lemma 3.6 when $\lambda > 4$ in Assumption (2.1). Due to Lemma 3.10 the leading term is $Q(Y_{t-1})$, provided Assumption (2.2) holds. This is $\mathrm{O_P}(1)$ by Lemma 3.11(ii) provided Assumptions (2.1), (2.2) are satisfied with $\lambda > 4$.

(ii) The term R_V is $\mathrm{O}(\log T)$ due to Lemma 3.5(i). The term R_ε vanishes due to Lemma 3.6 when $\lambda > 4$ in Assumption (2.1). Due to Lemma 3.8 the leading term is $\mathrm{O}(\log\log T)$.

(iii) Under (A) that $\mathbf{V}$ and $\mathbf{D}$ have no common eigenvalues then R_V is $\mathrm{O}\{(\log T)^{1/2}(\log\log T)^{1/2}\}$ due to Lemma 3.5(iii). The argument of (ii) can then be followed.

(iv) Under (B) that $\dim\mathbf{D} = 0$ with $\mathbf{V} = 1$ then R_V is $\mathrm{o}(1)$ due to Lemma 3.5(iv), whereas under (C) that $\dim\mathbf{V} = 0$ then $R_V = 0$. while it is $\mathrm{o}(1)$ under (B) $\dim\mathbf{D} = 0$ with $\mathbf{V} = 1$ due to Lemma 3.5(iv). The argument of (ii) can then be followed.

(v) The terms R_V and R_ε vanish as in (iv). As in (i) the leading term is $Q(Y_{t-1})$ by Lemma 3.10 provided Assumption (2.2) holds. This is of the desired order of magnitude by Lemma 3.11(i) provided Assumptions (2.2), (2.1) are satisfied with $\lambda > 4$. □

Remark 3.13. The condition $\lambda > 4$ in Theorem 2.5 can be relaxed as follows.

(i) It is used first in Lemma 3.6 and can be relaxed under (I,a) or (I,b) as this is a result holding in probability, see Remark 3.7. It is used secondly in Lemma 3.11(ii) and can be relaxed under (II), see Remark 3.12.

(ii), (iii), (iv) It is only used in Lemma 3.6 and can only be relaxed under (I,a) as this is a result holding almost surely, see Remark 3.7.

(v) It is indeed required in Lemma 3.11(i).

3.6. Proof of results for residual based tests

Proof. It suffices to show how the residual based test statistics relate to the likelihood ratio test statistics.

In the joint test the squared sample multiple correlation R^2 of $\hat{\varepsilon}_t$ and the vector $Z_{t-1} = (\hat{\varepsilon}'_{t-1}, \ldots, \hat{\varepsilon}'_{t-m}, \mathbf{X}'_{t-1}, D'_t)'$ is considered, recalling that $\mathbf{X}_{t-1}$ is defined as $(X'_{t-1}, \ldots, X'_{t-k+1})'$. The key to the result is that

$$\hat{\varepsilon}_{t-j} = X_{t-j} - \hat{\mathbf{B}}\mathbf{X}_{t-j-1} - \hat{\mu}\mathbf{D}^{-j-1}D_t,$$

where $\hat{\mathbf{B}}, \hat{\mu}$ are least squares estimators based on (1.1) for the full sample $t = 1, \ldots, T$. Due to the inclusion of $\mathbf{X}_{t-1}$ as regressor it follows that $Z_{t-1} = N\tilde{Z}_{t-1}$ where $\tilde{Z}_{t-1} = (X_{t-1}, \ldots, X_{t-k-m+1}, D_t)$ and the square matrix N is based on $\hat{\mathbf{B}}, \hat{\mu}$ and is invertible with probability one. By the invariance of sample multiple correlations to linear transformations then R^2 can be computed from $\hat{\varepsilon}_t$ and $\tilde{Z}_{t-1}$. By the same type of manipulation as in Lemma 3.2 it follows that

$$\hat{Q}\left(\tilde{Z}_{t-1}\right) = \sum_{t=m+1}^{T} \hat{\varepsilon}_t \tilde{Z}'_{t-1} \left(\sum_{t=m+1}^{T} \tilde{Z}_{t-1}^{\otimes 2} \right)^{-1} \sum_{t=m+1}^{T} \tilde{Z}_{t-1} \hat{\varepsilon}'_t$$

can be written as

$$\hat{Q}\left(\tilde{Z}_{t-1}\right) = \hat{Q}\left(X_{t-k}, \ldots, X_{t-k-m+1} | \mathbf{X}_{t-1}, D_t\right) + \hat{Q}\left(\mathbf{X}_{t-1}, D_t\right). \tag{3.13}$$

Since the first term in (3.13) includes the regressors $\mathbf{X}_{t-1}, D_t$ then $\hat{\varepsilon}_t$ can be replaced by ε_t. Thus, apart from starting the regression at $t = m + 1$ instead of $t = 1$ this term is the same as $Q(X_{t-k}, \ldots, X_{t-k-m+1} | \mathbf{X}_{t-1}, D_t)$. It therefore has the same asymptotic properties as $T(\hat{\Omega}_{k-1} - \hat{\Omega}_{k+m-1})$, which was studied in the proof of Theorem 2.3.

The second term in (3.13) vanishes asymptotically. This is because the residuals $\hat{\varepsilon}_t$ are orthogonal to $\mathbf{X}_{t-1}, D_t$ when evaluated over $t = 1, \ldots, T$. A tedious analysis shows that this orthogonality holds asymptotically when evaluated over $t = m + 1, \ldots, T$.

For the marginal test the argument is the same. The main difference is that the residuals are now

$$\hat{\varepsilon}_{t-j,\text{marg}} = X_{t-j,1} - \hat{\mathbf{B}}_{\text{marg}} \mathbf{X}_{t-j-1} - \hat{\mu}_{\text{marg}} \mathbf{D}^{-j-1} D_t.$$

Once again the inclusion of $\mathbf{X}_{t-1}$ as regressor implies that the vector $Z_{t-1,\text{marg}}$ defined as $(\hat{\varepsilon}'_{t-1}, \ldots, \hat{\varepsilon}'_{t-m}, \mathbf{X}'_{t-1}, D'_t)'$ can be replaced by the above $\tilde{Z}_{t-1}$. So the statistic $\hat{Q}(\tilde{Z}_{t-1})$ is replaced by a statistic based on $\hat{\varepsilon}_{t,\text{marg}}$, but the same $\tilde{Z}_{t-1}$.

For the conditional test the residuals are of the type

$$\hat{\varepsilon}_{t-j,\text{cond}} = X_{t-j,1} - \hat{\mathbf{B}}_{\text{cond}} \mathbf{X}_{t-j-1} - \hat{\mu}_{\text{cond}} \mathbf{D}^{-j-1} D_t - \hat{\omega} X_{t-j,2}.$$

The same argument applies as for the marginal test. □

Acknowledgments

Comments from the referee are gratefully acknowledged.

References

[1] Anderson, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Annals of Mathematical Statistics* **30** 676–687.

[2] Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ.

[3] Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers* **17** 334–355.

[4] Brown, B. M. and Eagleson, G. K. (1971). Martingale convergence to infinitely divisible laws with finite variance. *Transactions of the American Mathematical Society* **162** 449–453.
[5] Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* **16** 367–401.
[6] Donsker, M. D. and Varadhan, S. R. S. (1977). On laws of iterated logarithms for local times. *Communications on Pure and Applied Mathematics* **30** 707–753.
[7] Durbin, J. (1970). Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica* **38** 410–421.
[8] Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica* **46** 1293–1301.
[9] Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Annals of Statistics* **8** 1071–1081.
[10] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society* B **41** 190–195.
[11] Herstein, I. N. (1975). *Topics in Algebra*, 2nd edition. Wiley, New York.
[12] Heyde, C. C. and Scott, D. J. (1973). Invariance principles for the law of the iterated logarithm for martingales and processes with stationary increments. *Annals of Probability* **1** 428–436.
[13] Johansen, S. (2000). A Bartlett correction factor for tests on the cointegrating relations. *Econometric Theory* **16** 740–778.
[14] Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics* **10** 154–166.
[15] Lai, T. L. and Wei, C. Z. (1985). Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. In P. R. Krishnaiah (ed.), *Multivariate Analysis VI*, Elsevier Science Publishers, 375–393.
[16] Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.
[17] Nielsen, B. (2005). Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms. *Econometric Theory* **21** 534–561.
[18] Nielsen, B. (2006). Correlograms for non-stationary autoregressions. *Journal of the Royal Statistical Society*, B **68** 707–720.
[19] Pagan, A. R. (1984). Model evaluation by variable addition. In D. F. Hendry and K. F. Wallis (eds.), *Econometrics and Quantitative Economics*. Basil Blackwell, Oxford.
[20] Paulsen, J. (1984) Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* **5** 115–127.
[21] Pötscher, B. M. (1989). Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Annals of Statistics* **17** 1257–1274.
[22] Quinn, B. G. (1980). Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society* B **42** 182–185.
[23] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464.

[24] Tsay, R. S. (1984). Order selection in nonstationary autoregressive models. *Annals of Statistics* **12** 1425–1433.
[25] Wei, C. Z. (1992). On predictive least squares principles. *Annals of Statistics* **20** 1–42.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 113–129

DOI: 10.1214/074921706000000987

The distribution of model averaging estimators and an impossibility result regarding its estimation

Benedikt M. Pötscher[1]

University of Vienna

Abstract: The finite-sample as well as the asymptotic distribution of Leung and Barron's (2006) model averaging estimator are derived in the context of a linear regression model. An impossibility result regarding the estimation of the finite-sample distribution of the model averaging estimator is obtained.

1. Introduction

Model averaging or model mixing estimators have received increased interest in recent years; see, e.g., Yang [18–20], Magnus [13], Leung and Barron [12], and the references therein. [For a discussion of model averaging from a Bayesian perspective see Hoeting et al. [4].] The main idea behind this class of estimators is that averaging estimators obtained from different models should have the potential to achieve better overall risk performance when compared to a strategy that only uses the estimator obtained from one model. As a consequence, the above mentioned literature concentrates on studying the risk properties of model averaging estimators and on associated oracle inequalities. In this paper we derive the finite-sample as well as the asymptotic distribution (under fixed as well as under moving parameters) of the model averaging estimator studied in [12]; for the sake of simplicity we concentrate on the special case when only two candidate models are considered. Not too surprisingly, it turns out that the finite-sample distribution (after centering and scaling) depends on unknown parameters, and thus cannot be directly used for inferential purposes. As a consequence, one may be interested in estimators of this distribution, e.g., for purposes of conducting inference. We establish an impossibility result by showing that any estimator of the finite-sample distribution of the model averaging estimator is necessarily "bad" in a sense made precise in Section 4. While we concentrate on Leung and Barron's [12] estimator (in the context of only two candidate models) as a prototypical example of a model averaging estimator in this paper, similar results will typically hold for other model averaging estimators (and more than two candidate models) as well.

We note that results on distributional properties of post-model-selection estimators that parallel the development in the present paper have been obtained in [5–7, 9, 10, 14–17]. See also Leeb and Pötscher [11] for impossibility results pertaining to shrinkage-type estimators like the Lasso or Stein's estimator. An easily accessible exposition of the issues discussed in the just mentioned literature can be found in Leeb and Pötscher [8].

[1]Department of Statistics, University of Vienna, e-mail: benedikt.poetscher@univie.ac.at
AMS 2000 subject classifications: primary 62F10, 62F12; secondary 62E15, 62J05, 62J07.
Keywords and phrases: model mixing, model aggregation, combination of estimators, model selection, finite sample distribution, asymptotic distribution, estimation of distribution.

The only other paper we are aware of that considers distributional properties of model averaging estimators is Hjort and Claeskens [3]. Hjort and Claeskens [3] provide a result (Theorem 4.1) that says that – under some regularity conditions – the asymptotic distribution of a model averaging estimation scheme is the distribution of the same estimation scheme applied to the limiting experiment (which is a multivariate normal estimation problem). This result is an immediate consequence of the continuous mapping theorem, and furthermore becomes vacuous if the estimation problem one starts with is already a Gaussian problem (as is the case in the present paper).

2. The model averaging estimator and its finite-sample distribution

Consider the linear regression model

$$Y = X\beta + u$$

where Y is $n \times 1$ and where the $n \times k$ non-stochastic design matrix X has full column rank k, implying $n \geq k$. Furthermore, u is normally distributed $N(0, \sigma^2 I_n)$, $0 < \sigma^2 < \infty$. Although not explicitly shown in the notation, the elements of Y, X, and u may depend on sample size n. [In fact, the random variables Y and u may be defined on a sample space that varies with n.] Let $\mathbb{P}_{n,\beta,\sigma}$ denote the probability measure on $\mathbb{R}^n$ induced by Y, and let $\mathbb{E}_{n,\beta,\sigma}$ denote the corresponding expectation operator. As in [12], we also assume that σ^2 is known (and thus is fixed). [Results for the case of unknown σ^2 that parallel the results in the present paper can be obtained if σ^2 is replaced by the residual variance estimator derived from the unrestricted model. The key to such results is the observation that this variance estimator is independent of the least squares estimator for β. The same idea has been used in [7] to derive distributional properties of post-model-selection estimators in the unknown variance case from the known variance case. For brevity we do not give any details on the unknown variance case in this paper.] Suppose further that $k > 1$, and that X and β are commensurably partitioned as

$$X = [X_1 : X_2]$$

and $\beta = [\beta_1', \beta_2']'$ where X_i has dimension $k_i \geq 1$. Let the restricted model be defined as $M_R = \{\beta \in \mathbb{R}^k : \beta_2 = 0\}$ and let $M_U = \mathbb{R}^k$ denote the unrestricted model. Let $\hat{\beta}(R)$ denote the restricted least squares estimator, i.e., the $k \times 1$ vector given by

$$\hat{\beta}(R) = \begin{bmatrix} (X_1'X_1)^{-1}X_1'Y \\ 0_{k_2 \times 1} \end{bmatrix},$$

and let $\hat{\beta}(U) = (X'X)^{-1}X'Y$ denote the unrestricted least squares estimator. Leung and Barron [12] consider model averaging estimators in a linear regression framework allowing for more than two candidate models. Specializing their estimator to the present situation gives

$$\tilde{\beta} = \hat{\lambda}\hat{\beta}(R) + (1 - \hat{\lambda})\hat{\beta}(U) \tag{1}$$

where the weights are given by

$$\hat{\lambda} = [\exp(-\alpha\hat{r}(R)/\sigma^2) + \exp(-\alpha\hat{r}(U)/\sigma^2)]^{-1} \exp(-\alpha\hat{r}(R)/\sigma^2).$$

Here $\alpha > 0$ is a tuning parameter (note that Leung and Barron's tuning parameter corresponds to 2α) and

$$\hat{r}(R) = Y'Y - \hat{\beta}(R)'X'X\hat{\beta}(R) + \sigma^2(2k_1 - n)$$

and

$$\hat{r}(U) = Y'Y - \hat{\beta}(U)'X'X\hat{\beta}(U) + \sigma^2(2k - n).$$

For later use we note that

$$\begin{aligned}\hat{\lambda} &= [1 + \exp(-2\alpha k_2)\exp(-\alpha(\hat{\beta}(R)'X'X\hat{\beta}(R) - \hat{\beta}(U)'X'X\hat{\beta}(U))/\sigma^2)]^{-1} \\ &= [1 + \exp(-2\alpha k_2)\exp(\alpha \left\| X\hat{\beta}(R) - X\hat{\beta}(U)\right\|^2/\sigma^2)]^{-1}\end{aligned} \tag{2}$$

where $\|x\|$ denotes the Euclidean norm of a vector x, i.e., $\|x\| = (x'x)^{1/2}$. Leung and Barron [12] establish an oracle inequality for the risk $\mathbb{E}_{n,\beta,\sigma}(\|X(\tilde{\beta} - \beta)\|^2)$ and show that the model averaging estimator performs favourably in terms of this risk. As noted in the introduction, in the present paper we consider distributional properties of this estimator. Before we now turn to the finite-sample distribution of the model averaging estimator we introduce some notation: For a symmetric positive definite matrix A the unique symmetric positive definite root is denoted by $A^{1/2}$. The largest (smallest) eigenvalue of a matrix A is denoted by $\lambda_{\max}(A)$ ($\lambda_{\min}(A)$). Furthermore, P_R and P_U denote the projections on the column space of X_1 and of X, respectively.

Proposition 1. *The finite-sample distribution of $\sqrt{n}(\tilde{\beta} - \beta)$ is given by the distribution of*

$$\begin{aligned}&B_n\sqrt{n}\beta_2 + C_n\sqrt{n}Z_1 + \\ &\left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\left\|Z_2 + (X_2'(I - P_R)X_2)^{1/2}\beta_2\right\|^2/\sigma^2\right)\right]^{-1} \times \\ &\{D_n\sqrt{n}Z_2 - B_n\sqrt{n}\beta_2\}\end{aligned} \tag{3}$$

which can also be written as

$$\begin{aligned}&C_n\sqrt{n}Z_1 + D_n\sqrt{n}Z_2 - \\ &\left[1 + \exp(-2\alpha k_2)\exp\left(\alpha\left\|Z_2 + (X_2'(I - P_R)X_2)^{1/2}\beta_2\right\|^2/\sigma^2\right)\right]^{-1} \times \\ &\{D_n\sqrt{n}Z_2 - B_n\sqrt{n}\beta_2\}.\end{aligned} \tag{4}$$

Here

$$\begin{aligned}B_n &= \begin{bmatrix}(X_1'X_1)^{-1}X_1'X_2 \\ -I_{k_2}\end{bmatrix}, \qquad C_n = \begin{bmatrix}(X_1'X_1)^{-1/2} \\ 0_{k_2\times k_1}\end{bmatrix}, \\ D_n &= \begin{bmatrix}-(X_1'X_1)^{-1}X_1'X_2(X_2'(I - P_R)X_2)^{-1/2} \\ (X_2'(I - P_R)X_2)^{-1/2}\end{bmatrix},\end{aligned}$$

and Z_1 and Z_2 are independent, $Z_1 \sim N(0, \sigma^2 I_{k_1})$, and $Z_2 \sim N(0, \sigma^2 I_{k_2})$.

Proof. Observe that

$$\tilde{\beta} = \hat{\beta}(R) + (1 - \hat{\lambda})(\hat{\beta}(U) - \hat{\beta}(R)) = \hat{\beta}(R) + (1 - \hat{\lambda})(X'X)^{-1}X'(P_U - P_R)Y$$

with $P_R = X_1(X_1'X_1)^{-1}X_1'$ and $P_U = X(X'X)^{-1}X'$. Diagonalize the projection matrix $P_U - P_R$ as

$$P_U - P_R = \mathcal{U}\Delta\mathcal{U}'$$

where the orthogonal $n \times n$ matrix $\mathcal{U}$ is given by

$$\mathcal{U} = [\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3] = \left[X_1(X_1'X_1)^{-1/2} : (I - P_R)X_2(X_2'(I - P_R)X_2)^{-1/2} : \mathcal{U}_3\right]$$

with $\mathcal{U}_3$ representing an $n \times (n-k)$ matrix whose columns form an orthonormal basis of the orthogonal complement of the space spanned by the columns of X. The $n \times n$ matrix Δ is diagonal with the first k_1 as well as the last $n-k$ diagonal elements equal to zero, and the remaining k_2 diagonal elements being equal to 1. Furthermore, set $V = \mathcal{U}'Y$ which is distributed $N(\mathcal{U}'X\beta, \sigma^2 I_n)$. Then

$$\left\|X\hat{\beta}(U) - X\hat{\beta}(R)\right\|^2 = \|(P_U - P_R)Y\|^2 = \|\Delta V\|^2 = \|V_2\|^2$$

where V_2 is taken from the partition of $V' = (V_1', V_2', V_3')'$ into subvectors of dimensions k_1, k_2, and $n-k$, respectively. Note that V_2 is distributed $N((X_2'(I - P_R)X_2)^{1/2}\beta_2, \sigma^2 I_{k_2})$. Hence, in view of (2) we have that $(1-\hat{\lambda})(\hat{\beta}(U) - \hat{\beta}(R))$ is equal to

$$\begin{aligned}
&\left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|V_2\|^2/\sigma^2\right)\right]^{-1}(X'X)^{-1}X'\mathcal{U}\Delta V \\
&\quad = \left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|V_2\|^2/\sigma^2\right)\right]^{-1}(X'X)^{-1}\begin{bmatrix} 0_{k_1\times 1} \\ X_2'\mathcal{U}_2V_2 \end{bmatrix} \\
&\quad = \left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|V_2\|^2/\sigma^2\right)\right]^{-1} D_n V_2.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\hat{\beta}(R) &= (X'X)^{-1}X'P_RY \\
&= (X'X)^{-1}X'P_R\mathcal{U}V \\
&= (X'X)^{-1}X'X_1(X_1'X_1)^{-1/2}V_1 \\
&= \begin{bmatrix} (X_1'X_1)^{-1/2}V_1 \\ 0_{k_2\times 1} \end{bmatrix} = C_nV_1
\end{aligned}$$

with V_1 distributed $N((X_1'X_1)^{-1/2}X_1'X\beta, \sigma^2 I_{k_1})$. Hence, the finite sample distribution of $\tilde{\beta}$ is the distribution of

$$C_nV_1 + \left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|V_2\|^2/\sigma^2\right)\right]^{-1} D_nV_2 \tag{5}$$

where V_1 and V_2 are independent normally distributed with parameters given above. Defining Z_i as the centered versions of V_i, subtracting β, and scaling by $\sqrt{n}$ then delivers the result. □

Remark 2. (i) The first two terms in (3) represent the distribution of $\sqrt{n}(\hat{\beta}(R)-\beta)$, whereas the third term represents the distribution of $(1-\hat{\lambda})\sqrt{n}(\hat{\beta}(U) - \hat{\beta}(R))$. In (4), the first two terms represent the distribution of $\sqrt{n}(\hat{\beta}(U) - \beta)$, whereas the third term represents the distribution of $-\hat{\lambda}\sqrt{n}(\hat{\beta}(U) - \hat{\beta}(R))$.

(ii) If $\beta_2 = 0$ then (3) can be rewritten as

$$C_n\sqrt{n}Z_1 + \|Z_2\|\left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|Z_2\|^2/\sigma^2\right)\right]^{-1} D_n\sqrt{n}(Z_2/\|Z_2\|)$$

showing that this term has the same distribution as

$$C_n\sqrt{n}Z_1 + \sqrt{\chi^2}[1+\exp(2\alpha k_2)\exp\left(-\alpha\chi^2/\sigma^2\right)]^{-1}D_n\sqrt{n}U$$

where χ^2 is distributed as a χ^2 with k_2 degrees of freedom, $U = Z_2/\left\|Z_2\right\|$ is uniformly distributed on the unit sphere in $\mathbb{R}^{k_2}$, and Z_1, χ^2, and U are mutually independent.

Theorem 3. *The finite-sample distribution of $\sqrt{n}(\tilde{\beta}-\beta)$ possesses a density $f_{n,\beta,\sigma}$ given by*

$$\begin{aligned}
f_{n,\beta,\sigma}(t) &= (2\pi\sigma^2)^{-k/2}[\det(X'X/n)]^{1/2} \\
&\quad\times \exp\left(-(2\sigma^2)^{-1}\left\|n^{-1/2}(X_1'X_1)^{1/2}t_1 + n^{-1/2}(X_1'X_1)^{-1/2}X_1'X_2t_2\right\|^2\right) \\
&\quad\times \left[1+\exp\left(-\alpha\sigma^{-2}g\left(\left\|n^{-1/2}D_{n2}^{-1}(t_2+n^{1/2}\beta_2)\right\|\right)^2 + 2\alpha k_2\right)\right]^{k_2} \\
&\quad\times \left\{1+2\alpha\sigma^{-2}g\left(\left\|n^{-1/2}D_{n2}^{-1}(t_2+n^{1/2}\beta_2)\right\|\right)^2\right. \\
&\quad\times \left.\left[1+\exp\left(\alpha\sigma^{-2}g\left(\left\|n^{-1/2}D_{n2}^{-1}(t_2+n^{1/2}\beta_2)\right\|\right)^2 - 2\alpha k_2\right)\right]^{-1}\right\}^{-1} \\
&\quad\times \exp\left(-(2\sigma^2)^{-1}\left\|g\left(\left\|n^{-1/2}D_{n2}^{-1}(t_2+n^{1/2}\beta_2)\right\|\right)\right.\right. \\
&\quad\times \left.\left.\left\|n^{-1/2}D_{n2}^{-1}(t_2+n^{1/2}\beta_2)\right\|^{-1}n^{-1/2}D_{n2}^{-1}(t_2+n^{1/2}\beta_2) - D_{n2}^{-1}\beta_2)\right\|^2\right),
\end{aligned} \tag{6}$$

where t is partitioned as $(t_1', t_2')'$ with t_1 being a $k_1\times 1$ vector. Furthermore, $D_{n2} = (X_2'(I-P_R)X_2)^{-1/2}$, and g is as defined in the Appendix (with $a=\exp(2\alpha k_2)$ and $b=\alpha^{-1}\sigma^2$).

Proof. By (5) we have that the finite-sample distribution of $\sqrt{n}(\tilde{\beta}-\beta)$ is the distribution of

$$-\sqrt{n}\beta + \sqrt{n}[C_n : D_n][V_1' : V_3']'$$

where

$$V_3 = \left[1+\exp(2\alpha k_2)\exp\left(-\alpha\left\|V_2\right\|^2/\sigma^2\right)\right]^{-1}V_2.$$

By Lemmata 15 and 16 in the Appendix it follows that V_3 possesses the density

$$\begin{aligned}
\psi(v_3) &= (2\pi\sigma^2)^{-k_2/2}\left[1+\exp\left(-\alpha\sigma^{-2}g\left(\left\|v_3\right\|\right)^2 + 2\alpha k_2\right)\right]^{k_2} \\
&\quad\times\left\{1+2\alpha\sigma^{-2}g\left(\left\|v_3\right\|\right)^2\left[1+\exp\left(\alpha\sigma^{-2}g\left(\left\|v_3\right\|\right)^2 - 2\alpha k_2\right)\right]^{-1}\right\}^{-1} \\
&\quad\times\exp\left(-(2\sigma^2)^{-1}\left\|g\left(\left\|v_3\right\|\right)v_3/\left\|v_3\right\| - (X_2'(I-P_R)X_2)^{1/2}\beta_2\right\|^2\right).
\end{aligned}$$

Since V_1 is independent of V_2, and hence of V_3, the joint density of $[V_1' : V_3']'$ exists and is given by

$$(2\pi\sigma^2)^{-k_1/2}\exp\{-(2\sigma^2)^{-1}\left\|v_1 - (X_1'X_1)^{-1/2}X_1'X\beta\right\|^2\}\psi(v_3).$$

Since the matrix $[C_n : D_n]$ is non-singular we obtain for the density of $\sqrt{n}(\tilde{\beta}-\beta)$

$$
\begin{aligned}
&(2\pi\sigma^2)^{-k_1/2} n^{-k/2} [\det(X_1'X_1)\det(X_2'(I-P_R)X_2)]^{1/2} \\
&\times \exp\Big(-(2\sigma^2)^{-1} \Big\| n^{-1/2}(X_1'X_1)^{1/2}(t_1+n^{1/2}\beta_1) \\
&+ n^{-1/2}(X_1'X_1)^{-1/2}X_1'X_2(t_2+n^{1/2}\beta_2) - (X_1'X_1)^{-1/2}X_1'X\beta\Big\|^2\Big) \\
&\times \psi\Big(n^{-1/2}(X_2'(I-P_R)X_2)^{1/2}(t_2+n^{1/2}\beta_2)\Big).
\end{aligned}
$$

Note that $\det(X_1'X_1)\det(X_2'(I-P_R)X_2) = \det(X'X)$. Using this, and inserting the definition of ψ, delivers the final result (6). □

Remark 4. From Proposition 1 one can immediately obtain the finite-sample distribution of $\sqrt{n}A_n(\tilde{\beta}-\beta)$ by premultiplying (3) or (4) by A_n. Here A_n is an arbitrary (nonstochastic) $p_n \times k$ matrix. If A_n has full row-rank equal to k (implying $p_n = k$), this distribution has a density, which is given by $\det(A_n)^{-1} f_{n,\beta,\sigma}(A_n^{-1}s)$, $s \in \mathbb{R}^k$.

3. Asymptotic properties

For the asymptotic results we shall – besides the basic assumptions made in the preceding section – also assume that

$$\lim_{n\to\infty} X'X/n = Q \tag{7}$$

exists and is positive definite, i.e., $Q > 0$. We first establish "uniform $\sqrt{n}$-consistency" of the model averaging estimator, implying, in particular, uniform consistency of this estimator.

Theorem 5. *Suppose (7) holds.*

1. *Then $\tilde{\beta}$ is uniformly $\sqrt{n}$-consistent for β, in the sense that*

$$\lim_{M\to\infty} \sup_{n\geq k} \sup_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta,\sigma}\Big(\sqrt{n}\Big\|\tilde{\beta}-\beta\Big\| \geq M\Big) = 0. \tag{8}$$

   Consequently, for every $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta,\sigma}\Big(c_n\Big\|\tilde{\beta}-\beta\Big\| \geq \varepsilon\Big) = 0 \tag{9}$$

   holds for any sequence of real numbers $c_n \geq 0$ satisfying $c_n = o(n^{1/2})$; which reduces to uniform consistency for $c_n = 1$.
2. *The results in Part 1 also hold for $A_n\tilde{\beta}$ as an estimator of $A_n\beta$, where A_n are arbitrary (nonstochastic) matrices of dimension $p_n \times k$ such that the largest eigenvalues $\lambda_{\max}(A_n'A_n)$ are bounded.*

Proof. We prove (8) first. Rewrite the model averaging estimator as $\tilde{\beta} = \hat{\beta}(U) + \hat{\lambda}(\hat{\beta}(R) - \hat{\beta}(U))$. Since

$$\Big\|\tilde{\beta}-\beta\Big\| \leq \Big\|\hat{\beta}(U)-\beta\Big\| + \Big|\hat{\lambda}\Big|\,\Big\|\hat{\beta}(R)-\hat{\beta}(U)\Big\|,$$

since

$$\mathbb{P}_{n,\beta,\sigma}\Big(\sqrt{n}\Big\|\hat{\beta}(U)-\beta\Big\| \geq M\Big) \leq M^{-2}\sigma^2 \operatorname{trace}[(X'X/n)^{-1}],$$

and since $\text{trace}[(X'X/n)^{-1}] \to \text{trace}[Q^{-1}] < \infty$, it suffices to establish

$$\lim_{M\to\infty} \sup_{n\geq k} \sup_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta,\sigma}\left(\sqrt{n}\left|\hat{\lambda}\right|\left\|\hat{\beta}(R)-\hat{\beta}(U)\right\| \geq M\right) = 0. \tag{10}$$

Now, using (2) and the elementary inequality $z^2/[1+c\exp(z^2)]^2 \leq c^{-2}$ we have

$$\begin{aligned}
&\hat{\lambda}^2\left\|\hat{\beta}(R)-\hat{\beta}(U)\right\|^2 \\
&\quad\leq \hat{\lambda}^2\lambda_{\min}^{-1}(X'X)\left\|X\hat{\beta}(R)-X\hat{\beta}(U)\right\|^2 \\
&\quad= \lambda_{\min}^{-1}(X'X)\left[1+\exp(-2\alpha k_2)\exp\left(\alpha\left\|X\hat{\beta}(R)-X\hat{\beta}(U)\right\|^2/\sigma^2\right)\right]^{-2} \\
&\qquad\times\left\|X\hat{\beta}(R)-X\hat{\beta}(U)\right\|^2 \\
&\quad\leq n^{-1}\lambda_{\min}^{-1}(X'X/n)\alpha^{-1}\sigma^2\exp(4\alpha k_2) \leq Kn^{-1}\sigma^2
\end{aligned} \tag{11}$$

for a suitable finite constant K, since $\lambda_{\min}(X'X/n) \to \lambda_{\min}(Q) > 0$. This proves (10) and thus completes the proof of (8). The remaining claims in Part 1 follow now immediately. Part 2 is an immediate consequence of Part 1, of the inequality

$$\left\|A_n\tilde{\beta}-A_n\beta\right\|^2 \leq \lambda_{\max}(A_n'A_n)\left\|\tilde{\beta}-\beta\right\|^2,$$

and of the assumption on $\lambda_{\max}(A_n'A_n)$. □

Remark 6. (i) The proof has in fact shown that the difference between $\tilde{\beta}$ and $\hat{\beta}(U)$ is bounded in norm by a deterministic sequence of the form $const * \sigma n^{-1/2}$.

(ii) Although of little statistical significance since σ^2 is here assumed to be known, the proof also shows that the above proposition remains true if a supremum over $0 < \sigma^2 \leq S$, $(0 < S < \infty)$ is inserted in (8) and (9).

In the next two theorems we give the asymptotic distribution under general "moving parameter" asymptotics. Note that the case of fixed parameter asymptotics ($\beta^{(n)} \equiv \beta$) as well as the case of the usual local alternative asymptotics ($\beta^{(n)} = \beta + \delta/\sqrt{n}$) is covered by the subsequent theorems. In both these cases, Part 1 of the subsequent theorem applies if $\beta_2 \neq 0$, while Part 2 with $\gamma = 0$ and $\gamma = \delta_2$, respectively, applies if $\beta_2 = 0$.

Theorem 7. *Suppose (7) holds.*

1. *Let $\beta^{(n)}$ be a sequence of parameters such that $\|\sqrt{n}\beta_2^{(n)}\| \to \infty$ as $n \to \infty$. Then the distribution of $\sqrt{n}(\tilde{\beta}-\beta^{(n)})$ under $\mathbb{P}_{n,\beta^{(n)},\sigma}$ converges weakly to a $N(0,\sigma^2Q^{-1})$-distribution.*
2. *Let $\beta^{(n)}$ be a sequence of parameters such that $\sqrt{n}\beta_2^{(n)} \to \gamma \in \mathbb{R}^{k_2}$ as $n \to \infty$. Then the distribution of $\sqrt{n}(\tilde{\beta}-\beta^{(n)})$ under $\mathbb{P}_{n,\beta^{(n)},\sigma}$ converges weakly to the distribution of*

$$\begin{aligned}
&B_\infty\gamma + C_\infty Z_1 \\
&+\left[1+\exp(2\alpha k_2)\exp\left(-\alpha\left\|Z_2+(Q_{22}-Q_{21}Q_{11}^{-1}Q_{12})^{1/2}\gamma\right\|^2/\sigma^2\right)\right]^{-1} \\
&\times\{D_\infty Z_2 - B_\infty\gamma\}
\end{aligned} \tag{12}$$

where

$$B_\infty = \begin{bmatrix} Q_{11}^{-1}Q_{12} \\ -I_{k_2} \end{bmatrix}, \qquad C_\infty = \begin{bmatrix} Q_{11}^{-1/2} \\ 0_{k_2 \times k_1} \end{bmatrix},$$
$$D_\infty = \begin{bmatrix} -Q_{11}^{-1}Q_{12}(Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{-1/2} \\ (Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{-1/2} \end{bmatrix},$$

and where $Z_1 \sim N(0, \sigma^2 I_{k_1})$ *is independent of* $Z_2 \sim N(0, \sigma^2 I_{k_2})$. *The density of the distribution of (12) is given by*

$$\begin{aligned} f_{\infty,\gamma}(t) = {} & (2\pi\sigma^2)^{-k/2}[\det(Q)]^{1/2} \\ & \times \exp\left(-(2\sigma^2)^{-1}\left\|Q_{11}^{1/2}t_1 + Q_{11}^{-1/2}Q_{12}t_2\right\|^2\right) \\ & \times \left[1 + \exp\left(-\alpha\sigma^{-2}g\left(\left\|D_{\infty 2}^{-1}(t_2+\gamma)\right\|\right)^2 + 2\alpha k_2\right)\right]^{k_2} \\ & \times \left\{1 + 2\alpha\sigma^{-2}g\left(\left\|D_{\infty 2}^{-1}(t_2+\gamma)\right\|\right)^2\right. \\ & \left. \times \left[1 + \exp\left(\alpha\sigma^{-2}g\left(\left\|D_{\infty 2}^{-1}(t_2+\gamma)\right\|\right)^2 - 2\alpha k_2\right)\right]^{-1}\right\}^{-1} \\ & \times \exp\left\{-(2\sigma^2)^{-1}\left\|g\left(\left\|D_{\infty 2}^{-1}(t_2+\gamma)\right\|\right)\left\|D_{\infty 2}^{-1}(t_2+\gamma)\right\|^{-1}\right.\right. \\ & \left.\left. \times\ D_{\infty 2}^{-1}(t_2+\gamma) - D_{\infty 2}^{-1}\gamma\right\|^2\right\}, \end{aligned} \tag{13}$$

where t *is partitioned as* $(t_1', t_2')'$ *with* t_1 *being a* $k_1 \times 1$ *vector. Furthermore,* $D_{\infty 2} = (Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{-1/2}$, *and* g *is as defined in the Appendix (with* $a = \exp(2\alpha k_2)$ *and* $b = \alpha^{-1}\sigma^2$*).*

Proof. To prove Part 1 represent $\sqrt{n}(\tilde{\beta} - \beta^{(n)})$ as $\sqrt{n}(\hat{\beta}(U) - \beta^{(n)}) + \hat{\lambda}\sqrt{n}(\hat{\beta}(R) - \hat{\beta}(U))$. The first term is $N(0, \sigma^2(X'X/n)^{-1})$-distributed under $\mathbb{P}_{n,\beta^{(n)},\sigma}$, which obviously converges to a $N(0, \sigma^2 Q^{-1})$-distribution. It hence suffices to show that $\hat{\lambda}\sqrt{n}(\hat{\beta}(R) - \hat{\beta}(U))$ converges to zero in $\mathbb{P}_{n,\beta^{(n)},\sigma}$-probability. Since $\lambda_{\min}^{-1}(X'X/n)$ is bounded by assumption (7) and since

$$\begin{aligned} \hat{\lambda}^2\left\|\sqrt{n}(\hat{\beta}(R) - \hat{\beta}(U))\right\|^2 \le {} & n\lambda_{\min}^{-1}(X'X)\left\|X\hat{\beta}(R) - X\hat{\beta}(U)\right\|^2 \\ & \times \left[1 + \exp\left(\alpha\sigma^{-2}\left\|X\hat{\beta}(R) - X\hat{\beta}(U)\right\|^2 - 2\alpha k_2\right)\right]^{-2} \end{aligned}$$

as shown in (11), it furthermore suffices to show that

$$\left\|X\hat{\beta}(R) - X\hat{\beta}(U)\right\|^2 \to \infty \text{ in } \mathbb{P}_{n,\beta^{(n)},\sigma}\text{-probability.} \tag{14}$$

Note that

$$\begin{aligned} \left\|X\hat{\beta}(R) - X\hat{\beta}(U)\right\|^2 &= \|(P_U - P_R)Y\|^2 \\ &= \left\|(P_U - P_R)u + (P_U - P_R)X_2\beta_2^{(n)}\right\|^2 \\ &\ge \left|\left\|(P_U - P_R)X_2\beta_2^{(n)}\right\| - \|(P_U - P_R)u\|\right|^2. \end{aligned}$$

The second term satisfies $\mathbb{E}_{n,\beta^{(n)},\sigma}\|(P_U - P_R)u\|^2 = \sigma^2 k_2$ and hence is stochastically bounded in $\mathbb{P}_{n,\beta^{(n)},\sigma}$-probability. The square of the first term, i.e.,

$$\left\|(P_U - P_R)X_2\beta_2^{(n)}\right\|^2$$

equals

$$\sqrt{n}\beta_2^{(n)\prime}[(X_2'X_2/n) - (X_2'X_1/n)(X_1'X_1/n)^{-1}(X_1'X_2/n)]\sqrt{n}\beta_2^{(n)}.$$

Since the matrix in brackets converges to $Q_{22} - Q_{21}Q_{11}Q_{12}$, which is positive definite, the above display diverges to infinity, establishing (14). This completes the proof of Part 1.

We next turn to the proof of Part 2. The proof of (12) is immediate from (3) upon observing that $B_n \to B_\infty$, $\sqrt{n}C_n \to C_\infty$, and $\sqrt{n}D_n \to D_\infty$. To prove (13) observe that (12) can be written as

$$\begin{aligned}
&B_\infty\gamma + C_\infty Z_1 \\
&+ \left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\left\|Z_2 + (Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{1/2}\gamma\right\|^2/\sigma^2\right)\right]^{-1} \\
&\times \{D_\infty(Z_2 + (Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{1/2}\gamma)\} \\
&\quad = B_\infty\gamma + C_\infty Z_1 + D_\infty\left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|W_2\|^2/\sigma^2\right)\right]^{-1} W_2
\end{aligned}$$

where $W_2 \sim N((Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{1/2}\gamma, \sigma^2 I_{k_2})$ is independent of Z_1. Again using Lemmata 15 and 16 in the Appendix gives the density of

$$W_3 = \left[1 + \exp(2\alpha k_2)\exp\left(-\alpha\|W_2\|^2/\sigma^2\right)\right]^{-1} W_2$$

as

$$\begin{aligned}
\chi(w_3) = (2\pi\sigma^2)^{-k_2/2}&\left[1 + \exp\left(-\alpha\sigma^{-2}g(\|w_3\|)^2 + 2\alpha k_2\right)\right]^{k_2} \\
&\times\left\{1 + 2\alpha\sigma^{-2}g\left(\|w_3\|\right)^2\left[1 + \exp\left(\alpha\sigma^{-2}g(\|w_3\|)^2 - 2\alpha k_2\right)\right]^{-1}\right\}^{-1} \\
&\times\exp\left(-(2\sigma^2)^{-1}\left\|g\left(\|w_3\|\right)w_3/\|w_3\| - (Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{1/2}\gamma\right\|^2\right).
\end{aligned}$$

Since Z_1 is independent of Z_2, and hence of W_3, the joint density of $[Z_1' : W_3']'$ exists and is given by

$$(2\pi\sigma^2)^{-k_1/2}\exp\left(-(2\sigma^2)^{-1}\|z_1\|^2\right)\chi(w_3).$$

Since the matrix $[C_\infty : D_\infty]$ is non-singular we obtain finally

$$\begin{aligned}
&(2\pi\sigma^2)^{-k_1/2}\left[\det(Q_{11})\det(Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})\right]^{1/2} \\
&\times\exp\left(-(2\sigma^2)^{-1}\left\|Q_{11}^{1/2}(t_1 - Q_{11}^{-1}Q_{12}\gamma) + Q_{11}^{-1/2}Q_{12}(t_2 + \gamma)\right\|^2\right) \\
&\times\chi\left((Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{1/2}(t_2 + \gamma)\right).
\end{aligned}$$

Inserting the expression for χ derived above gives (13). □

Since in both cases considered in the above theorem the limiting distribution is continuous, the finite-sample cumulative distribution function (cdf)

$$F_{n,\beta^{(n)},\sigma}(t) = \mathbb{P}_{n,\beta^{(n)},\sigma}\left(\sqrt{n}(\tilde{\beta} - \beta^{(n)}) \leq t\right)$$

converges to the cdf of the corresponding limiting distribution even in the sup-norm as a consequence of the multivariate version of Polya's Theorem (cf. [1], Ex.6, [2]). We next show that the convergence occurs in an even stronger sense. Let f_∞ denote the density of the asymptotic distribution of $\sqrt{n}(\tilde{\beta} - \beta^{(n)})$ given in the previous theorem. That is, f_∞ is equal to $f_{\infty,\gamma}$ given in (13) if $\sqrt{n}\beta_2^{(n)} \to \gamma \in \mathbb{R}^{k_2}$, and is equal to the density of an $N(0, \sigma^2 Q^{-1})$-distribution if $\|\sqrt{n}\beta_2^{(n)}\| \to \infty$. For obvious reasons and for convenience we shall denote the $N(0, \sigma^2 Q^{-1})$-density by $f_{\infty,\infty}$.

Theorem 8. *Suppose the assumptions of Theorem 7 hold. Then the finite-sample density $f_{n,\beta^{(n)},\sigma}$ of $\sqrt{n}(\tilde{\beta} - \beta^{(n)})$ converges to f_∞, the density of the corresponding asymptotic distribution, in the L^1-sense. Consequently, the finite-sample cdf $F_{n,\beta^{(n)},\sigma}$ converges to the corresponding asymptotic cdf in total variation distance.*

Proof. In the case where $\sqrt{n}\beta_2^{(n)} \to \gamma \in \mathbb{R}^{k_2}$, inspection of (6), and noting that g as well as T^{-1} given in Lemma 15 are continuous, shows that (6) converges to (13) pointwise. In the case where $\|\sqrt{n}\beta_2^{(n)}\| \to \infty$, Lemma 17 in the Appendix and inspection of (6) show that (6) converges pointwise to the density of a $N(0, \sigma^2 Q^{-1})$-distribution. Observing that $f_{n,\beta^{(n)},\sigma}$ as well as f_∞ are probability densities, the proof is then completed by an application of Scheffé's lemma. □

Remark 9. We note for later use that inspection of (13) combined with Lemma 17 in the Appendix shows that for $\|\gamma\| \to \infty$ we have $f_{\infty,\gamma} \to f_{\infty,\infty}$ (the $N(0, \sigma^2 Q^{-1})$-density) pointwise on $\mathbb{R}^k$, and hence also in the L^1-sense. As a consequence, the corresponding cdfs converge in the total variation sense to the cdf of a $N(0, \sigma^2 Q^{-1})$-distribution.

Remark 10. The results in this section imply that the convergence of the finite-sample cdf to the asymptotic cdf does not occur uniformly w.r.t. the parameter β. [Cf. also the first step in the proof of Theorem 13 below.]

Remark 11. Theorems 7 and 8 in fact provide a characterization of all accumulation points of the finite sample distribution $F_{n,\beta^{(n)},\sigma}$ (w.r.t. the total variation topology) for arbitrary sequences $\beta^{(n)}$. This follows from a simple subsequence argument applied to $\sqrt{n}\beta_2^{(n)}$ and observing that $(\mathbb{R} \cup \{-\infty, \infty\})^{k_2}$ is compact; cf. also Remark 4.4 in [7].

Remark 12. Part 1 of Theorem 7 as well as the representation (12) immediately generalize to $\sqrt{n}A(\tilde{\beta} - \beta)$ with A a non-stochastic $p \times k$ matrix. If A has full row-rank equal to k, the resulting asymptotic distribution has a density, which is given by $\det(A)^{-1} f_\infty(A^{-1}s)$, $s \in \mathbb{R}^k$.

4. Estimation of the finite-sample distribution: an impossibility result

As can be seen from Theorem 3, the finite-sample distribution depends on the unknown parameter β, even after centering at β. Hence, it is obviously of interest to estimate this distribution, e.g., for purposes of conducting inference. It is easy to construct a consistent estimator of the cumulative distribution function $F_{n,\beta,\sigma}$ of the scaled and centered model averaging estimator $\tilde{\beta}$, i.e., of

$$F_{n,\beta,\sigma}(t) = \mathbb{P}_{n,\beta,\sigma}\left(\sqrt{n}(\tilde{\beta} - \beta) \leq t\right).$$

To this end, let $\hat{M}$ be an estimator that consistently decides between the restricted model M_R and the unrestricted model M_U, i.e., $\lim_{n\to\infty} \mathbb{P}_{n,\beta,\sigma}(\hat{M} = M_R) = 1$ if $\beta_2 = 0$ and $\lim_{n\to\infty} \mathbb{P}_{n,\beta,\sigma}(\hat{M} = M_U) = 1$ if $\beta_2 \neq 0$. [Such a procedure is easily constructed, e.g., from BIC or from a t-test for the hypothesis $\beta_2 = 0$ with critical value that diverges to infinity at a rate slower than $n^{1/2}$.] Define $\check{f}_n$ equal to $f^{\dagger}_{\infty,\infty}$, the density of the $N(0, \sigma^2(X'X/n)^{-1})$-distribution, on the event $\hat{M} = M_U$, and define $\check{f}_n$ equal to $f^{\dagger}_{\infty,0}$ otherwise, where $f^{\dagger}_{\infty,0}$ follows the same formula as $f_{\infty,0}$, with the only exception that Q is replaced by $X'X/n$. Then – as is proved in the Appendix –

$$\int_{\mathbb{R}^k} \left|\check{f}_n(z) - f_{n,\beta,\sigma}(z)\right| dz \to 0 \tag{15}$$

in $\mathbb{P}_{n,\beta,\sigma}$-probability as $n \to \infty$ for every $\beta \in \mathbb{R}^k$. Define $\check{F}_n$ as the cdf corresponding to $\check{f}_n$. Then for every $\delta > 0$

$$\mathbb{P}_{n,\beta,\sigma}\left(\left\|\check{F}_n - F_{n,\beta,\sigma}\right\|_{TV} > \delta\right) \to 0$$

as $n \to \infty$, where $\|\cdot\|_{TV}$ denotes the total variation norm. This shows that $\check{F}_n$ is a consistent estimator of $F_{n,\beta,\sigma}$ in the total variation distance. A fortiori then also

$$\mathbb{P}_{n,\beta,\sigma}\left(\sup_t \left|\check{F}_n(t) - F_{n,\beta,\sigma}(t)\right| > \delta\right) \to 0$$

holds.

The estimator $\check{F}_n$ just constructed has been obtained from the asymptotic cdf by replacing unknown quantities with suitable estimators. As noted in Remark 10, the convergence of the finite-sample cdf to their asymptotic counterpart does not occur uniformly w.r.t. the parameter β. Hence, it is to be expected that $\check{F}_n$ will inherit this deficiency, i.e., $\check{F}_n$ will not be uniformly consistent. Of course, this makes it problematic to base inference on $\check{F}_n$, as then there is no guarantee – at *any* sample size – that $\check{F}_n$ will be close to the true cdf. This naturally raises the question if estimators other than $\check{F}_n$ exist that are uniformly consistent. The answer turns out to be negative as we show in the next theorem. In fact, uniform consistency fails dramatically, cf. (17) below. This result further shows that uniform consistency already fails over certain shrinking balls in the parameter space (and thus a fortiori fails in general over compact subsets of the parameter space), and fails even if one considers the easier estimation problem of estimating $F_{n,\beta,\sigma}$ only at a given value of the argument t rather than estimating the entire function $F_{n,\beta,\sigma}$ (and measuring loss in a norm like the total variation norm or the sup-norm). Although of little statistical significance, we note that a similar result can be obtained for the problem of estimating the asymptotic cdf. Related impossibility results for post-model-selection estimators as well as for certain shrinkage-type estimators are given in [9–11].

In the result to follow we shall consider estimators of $F_{n,\beta,\sigma}(t)$ at a *fixed* value of the argument t. An estimator of $F_{n,\beta,\sigma}(t)$ is now nothing else than a real-valued random variable $\Gamma_n = \Gamma_n(Y, X)$. For mnemonic reasons we shall, however, use the symbol $\hat{F}_n(t)$ instead of Γ_n to denote an arbitrary estimator of $F_{n,\beta,\sigma}(t)$. This notation should not be taken as implying that the estimator is obtained by evaluating

an estimated cdf at the argument t, or that it is constrained to lie between zero and one. For simplicity, we give the impossibility result only in the simple situation where $k_2 = 1$ and Q is block-diagonal, i.e., X_1 and X_2 are asymptotically orthogonal. There is no reason to believe that the non-uniformity problem will disappear in more complicated situations.

Theorem 13. *Suppose (7) holds. Suppose further that $k_2 = 1$ and that Q is block-diagonal, i.e., the $k_1 \times k_2$ matrix Q_{12} is equal to zero. Then the following holds for every $\beta \in M_R$ and every $t \in \mathbb{R}^k$: There exist $\delta_0 > 0$ and ρ_0, $0 < \rho_0 < \infty$, such that any estimator $\hat{F}_n(t)$ of $F_{n,\beta,\sigma}(t)$ satisfying*

$$\mathbb{P}_{n,\beta,\sigma}\left(\left|\hat{F}_n(t) - F_{n,\beta,\sigma}(t)\right| > \delta\right) \overset{n\to\infty}{\longrightarrow} 0 \tag{16}$$

for every $\delta > 0$ (in particular, every estimator that is consistent) also satisfies

$$\sup_{\substack{\vartheta\in\mathbb{R}^k \\ ||\vartheta-\beta||<\rho_0/\sqrt{n}}} \mathbb{P}_{n,\vartheta,\sigma}\left(\left|\hat{F}_n(t) - F_{n,\vartheta,\sigma}(t)\right| > \delta_0\right) \overset{n\to\infty}{\longrightarrow} 1. \tag{17}$$

The constants δ_0 and ρ_0 may be chosen in such a way that they depend only on t, Q, σ, and the tuning parameter α. Moreover,

$$\liminf_{n\to\infty} \inf_{\hat{F}_n(t)} \sup_{\substack{\vartheta\in\mathbb{R}^k \\ ||\vartheta-\beta||<\rho_0/\sqrt{n}}} \mathbb{P}_{n,\vartheta,\sigma}\left(\left|\hat{F}_n(t) - F_{n,\vartheta,\sigma}(t)\right| > \delta_0\right) > 0 \tag{18}$$

and

$$\sup_{\delta>0} \liminf_{n\to\infty} \inf_{\hat{F}_n(t)} \sup_{\substack{\vartheta\in\mathbb{R}^k \\ ||\vartheta-\beta||<\rho_0/\sqrt{n}}} \mathbb{P}_{n,\vartheta,\sigma}\left(\left|\hat{F}_n(t) - F_{n,\vartheta,\sigma}(t)\right| > \delta\right) \geq \frac{1}{2}, \tag{19}$$

where the infima in (18) and (19) extend over all *estimators $\hat{F}_n(t)$ of $F_{n,\beta,\sigma}(t)$.*

Proof. Step 1: Let $\beta \in M_R$ and $t \in \mathbb{R}^k$ be given. Observe that by Theorems 7 and 8 the limit

$$F_{\infty,\gamma}(t) := \lim F_{n,\beta+(\eta,\gamma)'/\sqrt{n},\sigma}(t)$$

exists for every $\eta \in \mathbb{R}^{k_1}$, $\gamma \in \mathbb{R}^{k_2} = \mathbb{R}$, and does not depend on η. We now show that $F_{\infty,\gamma}(t)$ is non-constant in $\gamma \in \mathbb{R}$. First, observe that by Remark 9 and the block-diagonality assumption on Q

$$\lim_{\|\gamma\|\to\infty} F_{\infty,\gamma}(t) = \mathbb{P}\left(Q_{11}^{-1/2} Z_1 \leq t_1\right) \mathbb{P}\left(Q_{22}^{-1/2} Z_2 \leq t_2\right)$$

where Z_1 and Z_2 are as in Theorem 7, t is partitioned as $(t_1', t_2)'$ with t_2 a scalar, and $\mathbb{P}$ is the probability measure governing $(Z_1', Z_2)'$. Second, we have from (12) and the block-diagonality assumption on Q that $F_{\infty,\gamma}(t)$ is the product of

$$\mathbb{P}\left(Q_{11}^{-1/2} Z_1 \leq t_1\right)$$

with

$$\begin{aligned} \mathbb{P}\Bigg(&\left[1 + \exp(2\alpha)\exp\left(-\alpha\left(Z_2 + Q_{22}^{1/2}\gamma\right)^2/\sigma^2\right)\right]^{-1} \\ &\times \left(Q_{22}^{-1/2} Z_2 + \gamma\right) - \gamma \leq t_2\Bigg). \end{aligned} \tag{20}$$

Since $\mathbb{P}(Q_{11}^{-1/2}Z_1 \leq t_1)$ is positive and independent of γ, it suffices to show that (20) differs from $\mathbb{P}(Q_{22}^{-1/2}Z_2 \leq t_2)$ for at least one $\gamma \in \mathbf{R}$. Suppose first that $t_2 > 0$. Then specializing to the case $\gamma = 0$ in (20) it suffices to show that

$$\mathbb{P}\left(\left[1 + \exp(2\alpha)\exp\left(-\alpha Z_2^2/\sigma^2\right)\right]^{-1} Q_{22}^{-1/2} Z_2 \leq t_2\right). \tag{21}$$

differs from $\mathbb{P}(Q_{22}^{-1/2}Z_2 \leq t_2)$. But this follows from

$$\begin{aligned}
&\mathbb{P}\left(\left[1 + \exp(2\alpha)\exp\left(-\alpha Z_2^2/\sigma^2\right)\right]^{-1} Q_{22}^{-1/2} Z_2 \leq t_2\right) \\
&\quad = 1/2 + \mathbb{P}\left(Z_2 \geq 0, h(Z_2) \leq Q_{22}^{1/2} t_2\right) \\
&\quad = 1/2 + \mathbb{P}\left(0 \leq Z_2 \leq g\left(Q_{22}^{1/2} t_2\right)\right) \\
&\quad > 1/2 + \mathbb{P}\left(0 \leq Z_2 \leq Q_{22}^{1/2} t_2\right) \\
&\quad = \mathbb{P}\left(Q_{22}^{-1/2} Z_2 \leq t_2\right)
\end{aligned}$$

since h as defined in the Appendix (with $a = \exp(2\alpha)$ and $b = \sigma^2/\alpha$) is strictly monotonically increasing and satisfies $h(x) < x$ for every $x > 0$, which entails $g(y) > y$ for every $y > 0$. For symmetry reasons a dual statement holds for $t_2 < 0$. It remains to consider the case $t_2 = 0$. In this case (20) equals

$$\begin{aligned}
&\mathbb{P}\left(\left[1 + \exp(2\alpha)\exp\left(-\alpha\left(Z_2 + Q_{22}^{1/2}\gamma\right)^2/\sigma^2\right)\right]^{-1}\right. \\
&\qquad \left. \times \left(Z_2 + Q_{22}^{1/2}\gamma\right) \leq Q_{22}^{1/2}\gamma\right).
\end{aligned} \tag{22}$$

Let $\gamma > 0$ be arbitrary. Then (22) equals

$$\mathbb{P}\left(Z_2 + Q_{22}^{1/2}\gamma < 0\right) + \mathbb{P}\left(Z_2 + Q_{22}^{1/2}\gamma \geq 0, h\left(Z_2 + Q_{22}^{1/2}\gamma\right) \leq Q_{22}^{1/2}\gamma\right).$$

Arguing as before, this can be written as

$$\begin{aligned}
&\mathbb{P}\left(Z_2 + Q_{22}^{1/2}\gamma < 0\right) + \mathbb{P}\left(0 \leq Z_2 + Q_{22}^{1/2}\gamma \leq g\left(Q_{22}^{1/2}\gamma\right)\right) \\
&\quad > \mathbb{P}\left(Z_2 + Q_{22}^{1/2}\gamma < 0\right) + \mathbb{P}\left(0 \leq Z_2 + Q_{22}^{1/2}\gamma \leq Q_{22}^{1/2}\gamma\right) \\
&\quad = \mathbb{P}\left(Z_2 \leq 0\right) = \mathbb{P}\left(Q_{22}^{-1/2} Z_2 \leq 0\right)
\end{aligned}$$

which completes the proof of Step 1.

Step 2: We prove (17) and (18) first. For this purpose we make use of Lemma 3.1 in Leeb and Pötscher [11] with the notational identification $\alpha = \beta \in M_R$, $B = \mathbb{R}^k$, $B_n = \{\vartheta \in \mathbb{R}^k : \|\vartheta - \beta\| < \rho_0 n^{-1/2}\}$, $\varphi_n(\cdot) = F_{n,\cdot,\sigma}(t)$, and $\hat{\varphi}_n = \hat{F}_n(t)$, where ρ_0 will be chosen shortly. The contiguity assumption of this lemma is obviously satisfied; cf. also Lemma A.1 in [11]. It hence remains to show that there exists a value of ρ_0, $0 < \rho_0 < \infty$, such that δ^* defined in Lemma 3.1 of Leeb and Pötscher [11], which represents the limit inferior of the oscillation of $F_{n,\cdot,\sigma}(t)$ over B_n, is positive. Applying Lemma 3.5(a) of Leeb and Pötscher [11] with $\zeta_n = \rho_0 n^{-1/2}$ and the set G_0 equal to $G = \{(\eta', \gamma)' \in \mathbb{R}^k : \|(\eta', \gamma)'\| < 1\}$, it suffices to show that $F_{\infty,\gamma}(t)$ viewed as a function of $(\eta', \gamma)'$ is non-constant on the set $\{(\eta', \gamma)' \in \mathbb{R}^k :$

$\|(\eta', \gamma)'\| < \rho_0\}$; in view of Lemma 3.1 of Leeb and Pötscher [11], the corresponding δ_0 can then be chosen as any positive number less than one-half of the oscillation of $F_{\infty,\gamma}(t)$ over this set. That such a ρ_0 indeed exists now follows from Step 1. Furthermore, observe that $F_{\infty,\cdot}(t)$ depends only on α, Q, σ, and t. Hence, δ_0 and ρ_0 may be chosen such that they also only depend on these quantities. This completes the proof of (17) and (18).

To prove (19) we use Corollary 3.4 in [11] with the same identification of notation as above, with $\zeta_n = \rho_0 n^{-1/2}$, and with $V = \mathbb{R}^k$. The asymptotic uniform equicontinuity condition in that corollary is then satisfied in view of

$$\|\mathbb{P}_{n,\theta,\sigma} - \mathbb{P}_{n,\vartheta,\sigma}\|_{TV} \leq 2\Phi\left(\|\theta - \vartheta\| \, \lambda_{\max}^{1/2}(X'X)/(2\sigma)\right) - 1,$$

cf. Lemma A.1 in [11]. Given that the positivity of δ^* has already been established in the previous paragraph, applying Corollary 3.4 in [11] then establishes (19). □

Remark 14. The impossibility result given in the above theorem also holds for the class of randomized estimators (with $\mathbb{P}_{n,\cdot,\sigma}$ replaced by $\mathbb{P}^*_{n,\cdot,\sigma}$, the distribution of the randomized sample). This follows immediately from Lemma 3.6 in [11] and the attending discussion.

Appendix A: Some technical results

Let the function $h : [0,\infty) \to [0,\infty)$ be given by $h(\xi) = [1 + a\exp(-\xi^2/b)]^{-1}\xi$ where a and b are positive real numbers. It is easy to see that h is strictly monotonically increasing on $[0,\infty)$, is continuous, satisfies $h(0) = 0$ and $\lim_{\xi\to\infty} h(\xi) = \infty$. The inverse $g : [0,\infty) \to [0,\infty)$ of h clearly exists, is strictly monotonically increasing on $[0,\infty)$, is continuous, satisfies $g(0) = 0$ and $\lim_{\zeta\to\infty} g(\zeta) = \infty$. In the following lemma we shall use the natural convention that $g(\|y\|)y/\|y\| = 0$ for $y = 0$, which makes $y \to g(\|y\|)y/\|y\|$ a continuous function on all of $\mathbb{R}^m$.

Lemma 15. *Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be given by*

$$T(x) = \left[1 + a\exp(-\|x\|^2/b)\right]^{-1} x$$

where a and b are positive real numbers. Then T is a bijection. Its inverse is given by

$$T^{-1}(y) = g(\|y\|)y/\|y\|$$

where g has been defined above. Moreover, T^{-1} is continuously partially differentiable and $\|T^{-1}(y)\| = g(\|y\|)$ holds for all y.

Proof. If $y = 0$ it is obvious that $T(T^{-1}(y)) = 0 = y$ in view of the convention made above. Now suppose that $y \neq 0$. Then

$$\begin{aligned} T(T^{-1}(y)) &= [1 + a\exp\left(-g(\|y\|)^2/b\right)]^{-1} g(\|y\|)y/\|y\| \\ &= h\left(g(\|y\|)\right) y/\|y\| = y. \end{aligned}$$

Similarly, if $x = 0$ then $T^{-1}(T(x)) = 0$. Now suppose $x \neq 0$. Then $T(x) \neq 0$ and, observing that $\|T(x)\| = [1 + a\exp(-\|x\|^2/b)]^{-1}\|x\|$, we have

$$\begin{aligned} T^{-1}(T(x)) &= g\left(\|T(x)\|\right) T(x)/\|T(x)\| \\ &= g\left(\left[1 + a\exp\left(-\|x\|^2/b\right)\right]^{-1} \|x\|\right) x/\|x\| \\ &= g\left(h(\|x\|)\right) x/\|x\| = x. \end{aligned}$$

That T^{-1} is continuously partially differentiable follows from the corresponding property of T and the fact that the determinant of the derivative of T does never vanish as shown in the next lemma. The final claim is obvious in case $y \neq 0$, and follows from the convention made above and the fact that $g(0) = 0$ in case $y = 0$. □

Lemma 16. *Let T be as in the preceding lemma. Then the determinant of the derivative $D_x T$ is given by*

$$\left[1 + a \exp\left(-\|x\|^2 / b\right)\right]^{-m} \left\{1 + 2b^{-1} \left[1 + a^{-1} \exp\left(\|x\|^2 / b\right)\right]^{-1} \|x\|^2\right\}$$

which is always positive.

Proof. Elementary calculations show that

$$D_x T = \left[1 + a \exp\left(-\|x\|^2 / b\right)\right]^{-1} \times \left\{I_m + 2ab^{-1} \exp\left(-\|x\|^2 / b\right) \left[1 + a \exp\left(-\|x\|^2 / b\right)\right]^{-1} xx'\right\}.$$

Since the determinate of $I_m + cxx'$ equals $1 + cx'x$, the result follows. □

Lemma 17. *For g defined above we have*

$$\lim_{\zeta \to \infty} g(\zeta)/\zeta = 1$$

and

$$\lim_{\zeta \to \infty} \left((g(\zeta)/\zeta) - 1\right) \zeta = 0.$$

Proof. It suffices to prove the second claim:

$$\begin{aligned} \lim_{\zeta \to \infty} \left((g(\zeta)/\zeta) - 1\right) \zeta &= \lim_{\zeta \to \infty} \left(g(\zeta) - \zeta\right) = \lim_{\xi \to \infty} \left(g(h(\xi)) - h(\xi)\right) \\ &= \lim_{\xi \to \infty} \left(\xi - \left[1 + a \exp\left(-\xi^2/b\right)\right]^{-1} \xi\right) \\ &= \lim_{\xi \to \infty} \xi \left[1 + a^{-1} \exp\left(\xi^2/b\right)\right]^{-1} = 0. \end{aligned}$$

□

Proof (Verification of (15) in Section 5). In view of Theorem 8 it suffices to show that

$$\int_{\mathbb{R}^k} \left|\check{f}_n(z) - f_\infty(z)\right| dz \to 0$$

in $\mathbb{P}_{n,\beta,\sigma}$-probability as $n \to \infty$ for every $\beta \in \mathbb{R}^k$ where we recall that f_∞ is equal to $f_{\infty,\infty}$, the density of an $N(0, \sigma^2 Q^{-1})$-distribution, if $\beta_2 \neq 0$, and is equal to $f_{\infty,0}$

given in (13) if $\beta_2 = 0$. Now,

$$\begin{aligned}
&\mathbb{P}_{n,\beta,\sigma}\left(\int_{\mathbb{R}^k} \left|\check{f}_n(z) - f_\infty(z)\right| dz > \varepsilon\right) \\
&\quad = \mathbb{P}_{n,\beta,\sigma}\left(\int_{\mathbb{R}^k} \left|\check{f}_n(z) - f_\infty(z)\right| dz > \varepsilon, \hat{M} = M_R\right) \\
&\qquad + \mathbb{P}_{n,\beta,\sigma}\left(\int_{\mathbb{R}^k} \left|\check{f}_n(z) - f_\infty(z)\right| dz > \varepsilon, \hat{M} = M_U\right) \\
&\quad = \mathbb{P}_{n,\beta,\sigma}\left(\int_{\mathbb{R}^k} \left|f^\dagger_{\infty,0}(z) - f_\infty(z)\right| dz > \varepsilon, \hat{M} = M_R\right) \\
&\qquad + \mathbb{P}_{n,\beta,\sigma}\left(\int_{\mathbb{R}^k} \left|f^\dagger_{\infty,\infty}(z) - f_\infty(z)\right| dz > \varepsilon, \hat{M} = M_U\right)
\end{aligned}$$

where we have made use of the definition of $\check{f}_n$. If $\beta \in M_R$, then clearly the event $\hat{M} = M_U$ has probability approaching zero and hence the last probability in the above display converges to zero. Furthermore, if $\beta \in M_R$, the last but one probability reduces to

$$\mathbb{P}_{n,\beta,\sigma}\left(\int_{\mathbb{R}^k} \left|f^\dagger_{\infty,0}(z) - f_{\infty,0}(z)\right| dz > \varepsilon, \hat{M} = M_R\right)$$

which converges to zero since

$$\int_{\mathbb{R}^k} \left|f^\dagger_{\infty,0}(z) - f_{\infty,0}(z)\right| dz \to 0$$

in view of pointwise convergence of $f^\dagger_{\infty,0}$ to $f_{\infty,0}$ and Scheffé's lemma. [To be able to apply Scheffé's lemma we need to know that not only $f_{\infty,0}$ but also $f^\dagger_{\infty,0}(z)$ is a probability density. But this is obvious, as (13) defines a probability density for *any* symmetric and positive definite matrix Q.] The proof for the case where $\beta \in M_U$ is completely analogous noting that then $f_\infty = f_{\infty,\infty}$ holds. □

Acknowledgments

I would like to thank Hannes Leeb, Richard Nickl, and two anonymous referees for helpful comments on the paper.

References

[1] Billingsley, P. and Topsoe, F. (1967). Uniformity in weak convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **7** 1–16.

[2] Chandra, T. K. (1989). Multidimensional Polya's theorem. *Bulletin of the Calcutta Mathematical Society* **81** 227–231.

[3] Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98** 879–899.

[4] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial [with discussion]. *Statistical Science* **19** 382–417.

[5] Leeb, H. (2005). The distribution of a linear predictor after model selection: conditional finite-sample distributions and asymptotic approximations. *Journal of Statistical Planning and Inference* **134** 64–89.

[6] Leeb, H. (2006). The distribution of a linear predictor after model selection: unconditional finite-sample distributions and asymptotic approximations. *IMS Lecture Notes–Monograph Series* **49** 291–311.
[7] Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* **19** 100–142.
[8] Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory* **21** 21–59.
[9] Leeb, H. and Pötscher, B. M. (2005). Can one estimate the conditional distribution of post-model-selection estimators? Working Paper, Department of Statistics, University of Vienna. *Annals of Statistics* **34**, forthcoming.
[10] Leeb, H. and Pötscher, B. M. (2005). Can one estimate the unconditional distribution of post-model-selection estimators? Working Paper, Department of Statistics, University of Vienna.
[11] Leeb, H. and Pötscher, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk bound results. *Econometric Theory* **22** 69–97.
[12] Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52**, 3396–3410.
[13] Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *The Econometrics Journal* **5** 225–236.
[14] Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7** 163–185.
[15] Pötscher, B. M. and Novak, A. J. (1998). The distribution of estimators after model selection: large and small sample results. *Journal of Statistical Computation and Simulation* **60** 19–56.
[16] Sen, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* **7** 1019–1033.
[17] Sen P. K. and Saleh, A. K. M. E. (1987). On preliminary test and shrinkage M-estimation in linear models. *Annals of Statistics* **15** 1580–1592.
[18] Yang, Y. (2000). Combining different regression procedures for adaptive regression. *Journal of Multivariate Analysis* **74** 135–161.
[19] Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica* **13** 783–809.
[20] Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory* **20** 176–222.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 138–148

DOI: 10.1214/074921706000001003

Modeling macroeconomic time series via heavy tailed distributions

J. A. D. Aston[1]

Academia Sinica

Abstract: It has been shown that some macroeconomic time series, especially those where outliers could be present, can be well modelled using heavy tailed distributions for the noise components. Methods for deciding when and where heavy-tailed models should be preferred are investigated. These investigations primarily focus on automatic methods for model identification and selection. Current methods are extended to incorporate a non-Gaussian selection element, and various different criteria for deciding on which overall model should be used are examined.

1. Introduction

While time series analysis is a rich topic for theoretical research, the implications of such work can impact many applied sciences, including physical, biological and social. An example of the application of statistical time series methods in economics is the seasonal adjustment of macroeconomic data, one of the primary functions carried out by many statistical agencies worldwide. These adjustments allow comparisons of economic indicators to be made in the presence of seasonal variations, and allow economic decisions to be made without the confounding factors of seasonal fluctuations. The seasonally adjusted data is an unobserved component in the data, and must be estimated using a model, which can be either parametrically or non-parametrically specified. However, the estimates from the model can be seriously affected by changes in the data, especially outliers in the data.

Methods have been developed to account for outliers in commonly used time series models for seasonal adjustment using heavier tailed distributions than the Gaussian [3, 5, 10], such as the t-distribution or mixtures of normals. The aim of this paper is to present findings regarding how to select whether a heavy tailed model is required for a data set based on the performance of several model selection criteria.

Currently, most statistical agencies use one of two packages for seasonal adjustment, X-12-ARIMA [6] from the US Census Bureau, or TRAMO/SEATS [7] from the Bank of Spain. These two programs both seasonally adjust the data, but in intrinsically different ways. X-12-ARIMA uses prespecified filters to remove the seasonal component from the data, in a non-parametric fashion, while TRAMO/SEATS uses the ARIMA methodology of [4] to determine the model and estimate the seasonal component. It will be this second approach that will be generalised to include non-Gaussian components. Recently, a new integrated version

[1]Institute of Statistical Science, Academia Sinica, 128 Academia Sinica, Sec 2, Taipei 11529, Taiwan, ROC, Tel: +886-2-2783-5611 ext. 314 Fax: +886-2-2783-1523 e-mail: jaston@stat.sinica.edu.tw

AMS 2000 subject classifications: primary 91B82; secondary 62M10.

Keywords and phrases: seasonal adjustment, outliers, model selection, t-distribution, economic time series.

of the software has been released [11] and the methodology also applies to this package.

Two common types of outliers are additive outliers (shocks) and level shift (break) outliers. The first refers to a single data point that is out of character for the data given the model, whereas the second refers to a discontinuous jump either up or down in the level of the data at some point, and continuing on after that point for the rest of the series. These outliers lead to errors in the determination of the underlying components, and thus need to be accounted for. However, they are, by nature, not known prior to the modeling, and are functions of both the data and the model. A data point may be an outlier for one model but not another, and its status as an outlier may change when new data is added. This second feature is especially relevant for this study, as seasonal macroeconomic data is usually continuously updated month by month, or quarter by quarter.

Firstly, a very brief introduction to ARIMA models for seasonal adjustment will be given, followed by characterisations of the usual methods of outlier detection for both the Gaussian and non-Gaussian cases. While attempting to solve similar problems, the two approaches are appreciably different. Section 4 introduces the model selection criteria to be considered and also some justification for their usage. Section 5 provides some examples of real series where adjustment using heavy tailed models gives better performance than using Gaussian models and the final section provides discussion.

2. Seasonal adjustment and ARIMA model based decomposition

A seasonal time series y_t can be expressed as the sum of unobserved components,

$$y_t = S_t + N_t \tag{1}$$

where S_t represents the seasonal component and N_t, the remaining non-seasonal component. Thus the seasonally adjusted series $y^{(sa)}$ is

$$y_t^{(sa)} = y_t - S_t \tag{2}$$

and is also unobserved by the nature of being a function of S_t. ARIMA model based (AMB) decomposition specifies ARIMA models for the unobserved components and estimates these from both the data and from the overall model for the data as a whole.

Box and Jenkins [4] introduced a class of seasonal ARIMA models that model macroeconomic data well. The simplest model of this form is the airline model, which models differenced data as a product of moving average (MA) processes;

$$(1 - B)(1 - B^s)y_t = (1 - \theta B)(1 - \Theta B^s)\epsilon_t \tag{3}$$

where $By_t = y_{t-1}$, s represents the seasonal periodicity, and θ and Θ are the MA parameters associated with the non-seasonal and seasonal MA parts respectively. ϵ_t is assumed to be an iid Gaussian white noise process with variance σ^2. This model can be generalised by altering the degrees of the MA polynomials and adding Autoregressive (AR) parts to the left hand side of the equation. It will be assumed here that the differencing is not modified, and that it remains of an airline type. This includes the restriction that the MA and AR parameters cannot be unit roots as these would alter the overall differencing.

For simplicity, the airline model is considered explicitly, although other generalisations can be handled similarly. The AMB decomposition for the airline model, can be expressed in the following way using the decomposition of [9]

$$\begin{aligned} U(B)S_t &= \theta_S(B)\omega_t \\ (1-B)^2 T_t &= \theta_T(B)\eta_t \\ I_t &= \varepsilon_t \end{aligned} \tag{4}$$

where $U(B) = (1 + B + \cdots + B^{s-1})$ and $\omega_t, \eta_t, \epsilon_t$ are independent white noise processes and

$$y_t = S_t + T_t + I_t. \tag{5}$$

In order to define a unique solution (which may or may not exist), the restriction is taken that the pseudo-spectral densities of the seasonal and trend components have a minimum of zero (in line with the admissible decompositions of [9]). When this condition cannot be met without resulting in a negative variance for I_t, the decomposition is said to be inadmissible. Throughout the paper, only parameter combinations resulting in admissible decompositions will be assumed, which is almost always the case for macroeconomic data.

The parameter functions $\theta_T(), \theta_S()$ and the variances of $\omega_t, \eta_t, \epsilon_t$ are all functions of the underlying parameters θ, Θ and σ^2. They can be calculated from the partial fraction decomposition of the pseudo-spectral densities, and the minimisation of each resulting component. This is usually done, for example in the SEATS software, after maximum likelihood estimation of the parameters has taken place, to give a final adjustment of the data.

2.1. Gaussian outlier adjustment

The TRAMO package [7] is the most widely used method for automatic model identification of seasonal ARIMA models for macroeconomic series. The program is used to estimate the order of differencing, the orders of the AR and MA components and also any outliers and common regressor effects that might be present. Inherently in this paper, it has been assumed that the order of differencing is the same as the airline model, but this assumption can be easily relaxed without significant change in the approaches outlined. All the other parts of the TRAMO procedure are used in exactly the same way in this paper as given in [8] except for the part relating to outlier detection.

The TRAMO software determines outliers as part of the automatic model identification portion of the program. Critical values for the thresholds at which data points are assumed to be outliers are chosen either by the user or from the length of the series. Outliers are found by determining whether the significance of the regression coefficients determined by assuming an outlier, be it an additive outlier or a level shift, has occurred at each point in the data. This is done iteratively, by adding in the largest regressor above the threshold (if one exists) and then repeating the exercise. A final check is made at the end to ensure that all regressors are still above the threshold for the final model.

This method effectively removes the data point when it is considered to be an outlier. When new data points arrive every month/quarter, the stability of the seasonal adjustment can heavily rely on the stability of the designated outliers to this new data as an outlying data point can then be added back in to the estimation if no longer classified as an outlier.

3. Heavy-tailed models to account for outliers in ARIMA component models

Aston and Koopman [3] proposed this alternate methodology to the Gaussian classification of outliers by using heavier tailed distributions to weight data points rather than making binary decisions on outliers. A short summary of the methodology is now given.

The decomposition model (4) can be modified to incorporate non-Gaussian components. In order to retain a similar structure, it is assumed that the components have the same variance as the decomposition would predict, but different densities are used to incorporate heavier tails.

The irregular component can be modified to include the t distribution in order to account for additive outliers as

$$I_t^* \sim t(0, \sigma_I^2, \nu), \qquad t = 1, \ldots, n, \tag{6}$$

where $\nu > 2$ is the number of degrees of freedom and σ_I^2 is the variance, which is constant for any ν. In the case of an irregular modelled by a mixture of normals,

$$I_t^* \sim (1-\rho)\mathcal{N}(0, \sigma_I^2) + \rho\mathcal{N}(0, \sigma_I^2\lambda), \qquad t = 1, \ldots, n, \tag{7}$$

where $0 \leq \rho \leq 1$ determines the intensity of outliers in the series and λ measures the magnitude of the outliers.

The decomposition model with a t-distributed irregular term can be expressed in its canonical form by

$$y_t = S_t + T_t + I_t^*, \qquad I_t^* \sim t(0, \sigma_I^2, \nu), \qquad t = 1, \ldots, n.$$

where $t(0, \sigma_I^2, \nu)$ refers to the t-density. This model has the same number of parameters as the original model specification except that the t density has one additional parameter (the degrees of freedom ν) and the mixture of normals has two additional parameters (the intensity and the variance scalar).

To robustify the decomposition model against breaks in trend we consider the trend specification

$$(1-B)^2 T_t^* = \theta_T(B)\eta_t^*, \qquad \eta_t^* \sim t(0, \sigma_\eta^2, \nu_\eta), \tag{8}$$

where the t-distribution $t(0, \sigma_\eta^2, \nu_\eta)$ can be replaced by a mixture of normals distribution. The decomposition model with heavy tailed densities for both the trend innovations and the irregular is given by $y_t = S_t + T_t^* + I_t^*$ where the latter two components are given by (8) and (6), respectively.

These models can be estimated through the use of importance sampling as was described in [3]. For calculation, it is important to note that the decomposition must now be incorporated into the maximum likelihood estimation, as the individual components are modified, yet, they are still dependent on the overall ARIMA model for the series. However, fast algorithms for the decomposition make the estimation feasible.

4. Model selection

One of the most important issues is deciding which model to use and when. Here three different approaches are investigated, one based on the moments of the data,

another using an empirical evaluation of a model selection criteria and the last based on the stability of the estimated components when the data is updated.

Although the model is complex and the estimation method of the maximum likelihood an approximation, certain properties of the model can be usefully investigated by considering a simpler model. Take the simple case of choosing between two noise models, one t-distributed and the other gaussian,

$$\begin{aligned} y_\tau &= \epsilon_\tau \;\; \epsilon_\tau \sim N(0, \phi_N) \\ y_\tau &= \epsilon_\tau \;\; \epsilon_\tau \sim t(0, \phi_T, \nu). \end{aligned} \tag{9}$$

and all the y_τ are iid from one model or the other.

As can be seen, these are essentially nested models where the parameter of interest is the ν degrees of freedom parameter. Slightly different from the usual nesting setup, the gaussian model is the limiting distribution of the t-model as $\nu \to \infty$. A proof is given in the appendix to show by looking at a function of the moments of the data (essentially the kurtosis), a test can be performed as to whether the error term under investigation comes from a normal or t-distributed model. The test simply considers whether $\sqrt{n}(Z_n - 3)$ comes from $N(0, 24)$, as should be the case asymptotically, if the data are normally distributed, while for the t-distribution, the sequence will diverge. In the actual data case, the test will be applied to the irregular component data fitted under the normal model for parameter estimation.

As can be seen from the simulations in Figure 1, even for small samples of the size of the real data under investigation, there is a marked difference in the distribution of the statistic between the two models.

In addition to the model choice given above, two other methods are investigated. AIC [2] seems to be well suited to this problem, as the models are essentially nested. Howev likelihood

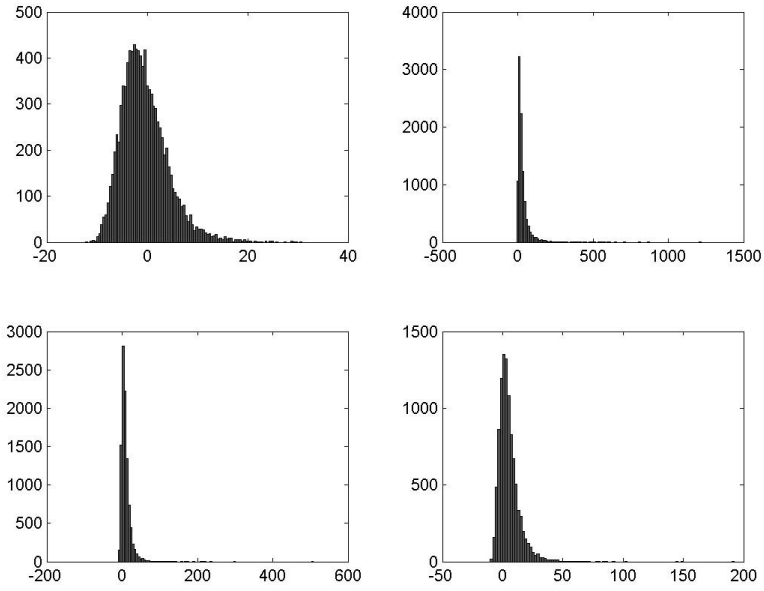
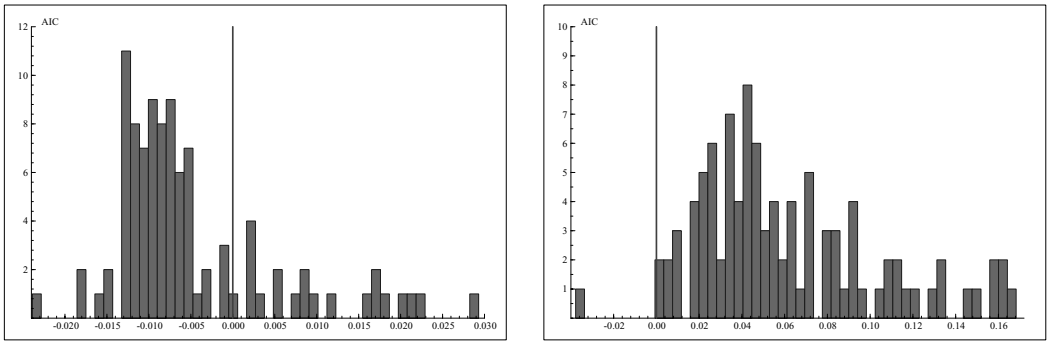


FIG 1. *Small sample kurtosis estimator distributions for four different models as generated from* 10^5 *simulated samples of* $n = 150$. *(top left) Gaussian, (top right) t dist* ($\nu = 5$), *(bottom left) t dist* ($\nu = 10$), *(bottom right) t dist* ($\nu = 15$)



(a) AIC: No Outliers (b) AIC: 3 Outliers

FIG 2. *AIC distributions for* 100 *series.* (a) *No outliers are present in the data* (b) *Three Outliers are present in the data. Black line corresponds to the value where AIC chooses Gaussian vs t-distributed model (left and right sides respectively).*

estimate of the parameters be an interior point within the parameter space, not on the boundary. Here, as the t distribution becomes the normal distribution as $\nu \to \infty$, this assumption is violated. However, in practice it can be seen that while the use of AIC might not yet be theoretically justified, in the case of the type of data under investigation, simulation results seem to be promising.

A small simulation study was carried out using the airline model (parameters of $\theta = 0.7$, $\Theta = 0.7$, $\sigma^2 = 1$) and two data sets generated, one where there were no outliers in the data, and one where there were three additive outliers added to the data (points shifted by 5 times the sd of the irregular component). Histograms of the AIC differences have been plotted in Figure 2. Corresponding histograms for AICc and BIC have also been generated (not shown) with similar results. AIC chooses the larger (t-distributed) model when there are outliers present in the data, and chooses the smaller model when outliers are not present, with an error rate of the same order as traditional AIC would predict.

In addition to the model selection criteria, an empirical measure was assessed for determining which model to use. This measure relies on using out-of-sample data to examine the estimates of the in-sample seasonally adjusted data when future data becomes available. A crude, yet seemingly promising, procedure is to withhold the final year of data, and to plot the changes in the seasonal components from the two samples, with and without the extra year of data. Given the problems of revisions when releasing macroeconomic data, adjusted series that remain stable when future data is added are to be preferred to adjusted series that change. By examining the plots of the differences, or some overall average change, such as the mean absolute difference between the two adjustments, the stability of the seasonal component to additional data can be quantified. Whilst this statistic is hard to justify theoretically given the complex nature of the model, it will be seen in the examples that it does seem to capture differences between the two approaches. Theoretical justification of this statistic will be the subject of future work.

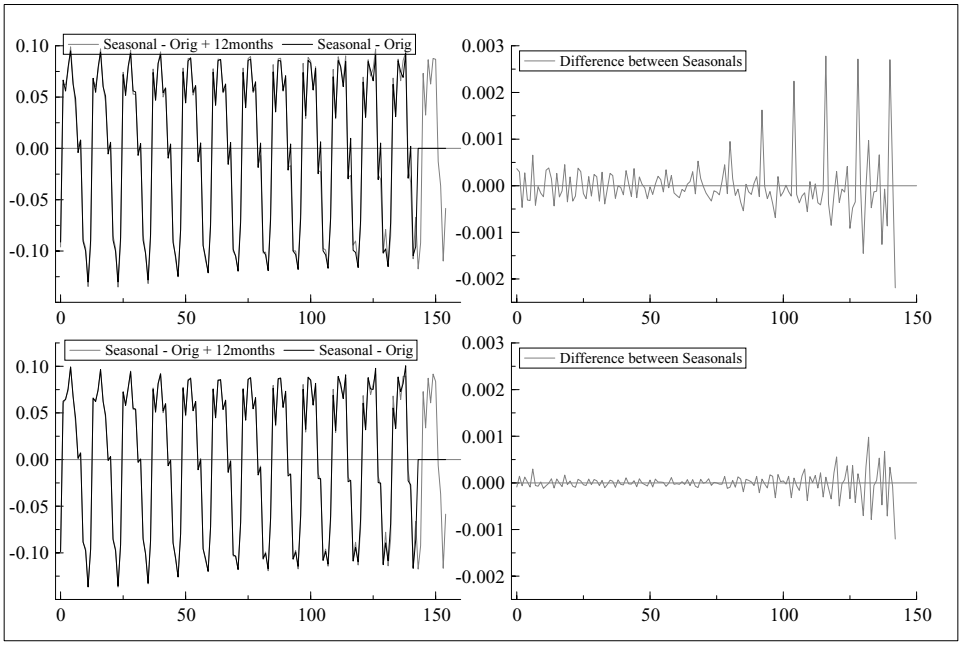
5. Examples

Many macroeconomic series do not have large problems with outliers, and thus the methods described here will not be applicable. However, there are a sizeable proportion of series released by agencies such as the US Census Bureau where outliers do occur. When several series from the Census Bureau were investigated, two series

where additive outliers seemed to be present were the Automobile Retail Series and the Material Handling Equipment Manufacturing Series (or u33mno as it is also known). These two series were analysed with the two different approaches, and the model selection criteria were used to determine which was the most appropriate model. Both series contained 155 data points (Feb 1992 until Dec 2004). For the seasonal stability plots, the data from 2004 was withheld in one analysis and used in the other and the results compared.

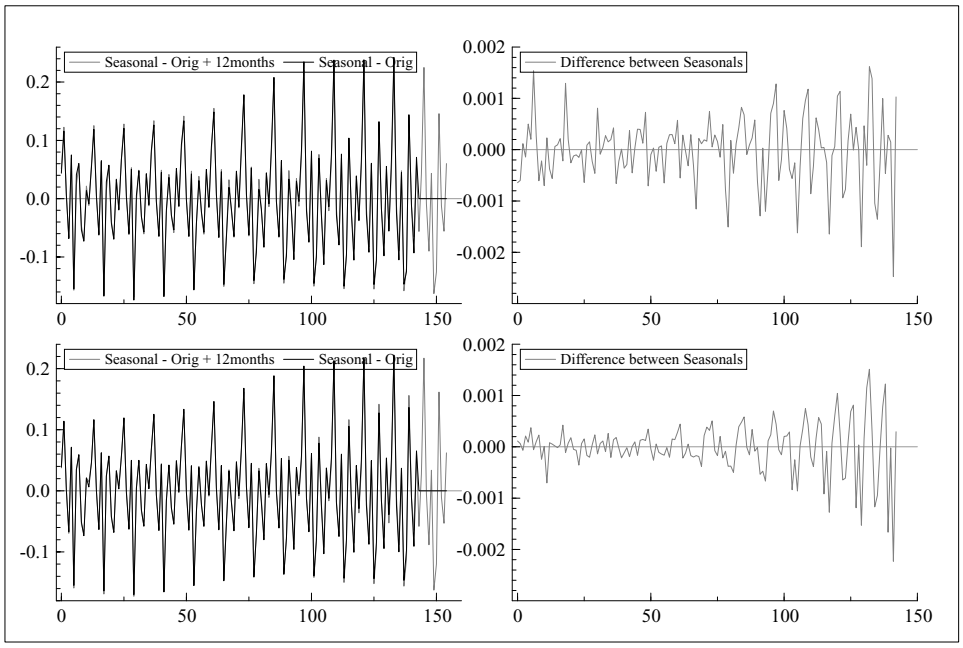
As can be seen in Figure 3, the seasonal stability plots for the automobile retail series indicate that the t-distribution provides a more stable adjustment than the normal model. This is chiefly because when an extra year of data is added, the number of outliers detected in the series changed. Thus this caused large changes in the seasonal pattern for the normal model. However, as there is no discrete detection process for the t-distributed model, there was a more continuous change in the seasonal pattern when the extra year of data was added, and thus the stability was greater. There was a low number of degrees of freedom (approximately 5-6) estimated for this model. The difference between the two models is also well detected by all the model selection criteria (Table 1). In addition, the moment estimator has a large value, well outside the 95% confidence interval range of a N(0,24) and therefore normality of the error terms are rejected for both the 143 and 155 length series.

The same conclusion can be reached with Figure 4 for the u33mno series, although the seasonal pattern was more erratic for both models and thus the stability of the seasonal was closer in both models. This was also shown in smaller differences for the model selection criteria in Table 1, with all the criteria being borderline as to which model to use, especially given the small sample nature of



Seasonal Differences (Log Data). top row: Gaussian Data, bottom row: t-distributed model data, left column: Seasonal Patterns, right column: RMS Differences in Patterns when extra 12 observations added

FIG 3. *Automobile Retail Series from Feb 1992-Dec 2004 (US Census Bureau). This example shows that the seasonal difference plot finds a large change in the seasonal pattern with an extra year of data when a Gaussian model is used, but this change is reduced when using the model containing the t-distribution*



Seasonal Differences (Log Data). top row: Gaussian Data, bottom row: t-distributed model data, left column: Seasonal Patterns, right column: RMS Differences in Patterns when extra 12 observations added

FIG 4. *u33mno Series from Feb 1992-Dec 2004 (US Census Bureau). Again, this example shows that the seasonal difference plot finds a large change in the seasonal pattern with an extra year of data when a Gaussian model is used, but this change is reduced (slightly less than the auto retail series) when using the model containing the t-distribution.*

TABLE 1
Model selection and comparison of the two example series. Bold face indicates the optimal value for selecting the model

Data	Model	Length	$\hat{df}$	Sample Kurtosis	LogLik	AIC	AICc	BIC	Seas Mean Abs Diff
Auto	G	143	-	35.8	217.18	-2.97	-1.95	-2.86	-
Auto	T	143	5.63	-	223.00	**-3.04**	**-2.02**	**-2.91**	-
Auto	G	155	-	47.5	236.75	-2.99	-1.97	-2.89	0.0629
Auto	T	155	5.10	-	243.05	**-3.06**	**-2.04**	**-2.94**	**0.0381**
u33mno	G	143	-	11.1	92.29	-1.22	-0.20	**-1.12**	-
u33mno	T	143	9.17	-	93.59	**-1.23**	**-0.21**	-1.10	-
u33mno	G	155	-	13.9	105.19	-1.40	-0.38	**-1.30**	0.0651
u33mno	T	155	7.43	-	107.14	**-1.41**	**-0.39**	-1.29	**0.0266**

series. However, given the increase in the stability and the borderline nature, the t-distribution model will probably be preferred in the case of u33mno as well.

It can be noted in both Figures 3 and 4 that, for both the t-distributed and the Gaussian models, the instability within the estimates does increase towards the end of the series. This is due to estimates being weighted functions of other data. Data that is close to the point to be estimated (either directly or as a multiple of the seasonal period) is more heavily weighted than data that is further away. Thus when new data is added, the estimates towards the end of the series are more heavily affected than the estimates nearer the beginning of the series.

6. Discussion

In this paper, model selection criteria have been proposed to choose between Gaussian and heavier tailed distributions. Particular emphasis has been placed on choosing between the t-distribution and the Gaussian distribution for modeling the irregular component where additive outliers occur. However, all the techniques are generalisable to other distributions such as mixtures of normals and to the trend component to account for level shifts.

The model selection criteria have primarily been evaluated empirically for the data and models used in the paper. This is for two reasons. Firstly, the seasonal models under investigation are complex models, where the likelihood evaluation involves both approximations through importance sampling and also pseudo-spectral decomposition. Thus, results for these types of models are difficult to obtain explicitly. However, even for simpler models, only theoretical results have been obtained for the moment estimator selection procedure, given the nature of the model nesting, and the boundary problem. However, the results obtained from simulation are promising. This also suggests that theoretical justification of these and related results, which apply in many other modeling situations, may well be a worthwhile future research area.

Acknowledgments

The author would like to express his gratitude to Ching-Kang Ing for all his help and suggestions with this work. He would also like to thank Siem Jan Koopman, Benedikt Pötscher and David Findley and an anonymous reviewer for their extremely helpful comments and discussions.

Appendix A: Moment estimator

Theorem 1 (Sample kurtosis). *Let y_τ, $1 \leq \tau \leq n$ be a realisation from one of the two models given in* (9) *with finite positive variance.*

Let

$$Z_n = \frac{\frac{1}{n}\sum (y_i - \bar{y})^4}{\left(\frac{1}{n}\sum (y_i - \bar{y})^2\right)^2} \tag{10}$$

and if $\rightsquigarrow$ represents convergence in distribution then

$$\sqrt{n}(Z_n - 3) \rightsquigarrow N(0, 24) \tag{11}$$

when $\nu \to \infty$ and diverges for ν finite.

Proof. Both the denominator and numerator of Z_n are moment estimators. If y_τ comes from the first model in (9) then following a similar method to [12, Example 3.5], let

$$\phi(a, b, c, d) = \frac{d - 4ca - 6ba^2 - 3a^4}{(b - a^2)^2}$$

then

$$Z_n = \phi(\overline{Y}, \overline{Y^2}, \overline{Y^3}, \overline{Y^4})$$

where $\overline{Y^j} = \frac{i=1}{n}\sum_1^n y_i^j$ and $\sqrt{n}(\overline{Y} - \alpha_1, \overline{Y^2} - \alpha_2, \overline{Y^3} - \alpha_3, \overline{Y^4} - \alpha_4)$ is asymptotically mean zero normal by the CLT where α_j is the jth moment of y_1 wlog.

If $X_i = \frac{y_i}{\phi_N}$, and using the fact that the odd moments of the standard normal are zero and the even moments given by $\frac{2n!}{2^n n!}$ (and thus the first eight moments are also finite),

$$\sqrt{n}\begin{pmatrix}\overline{X}\\ \overline{X^2}-1\\ \overline{X^3}\\ \overline{X^4}-3\end{pmatrix} \rightsquigarrow N\left(0,\begin{pmatrix}1&0&3&0\\0&2&0&12\\3&0&15&0\\0&12&0&96\end{pmatrix}\right).$$

The function ϕ is differentiable at the point $(\alpha_1,\alpha_2,\alpha_3,\alpha_4)=(0,1,0,3)$, and equals (0,-6,0,1). Hence, by use of the delta method

$$\sqrt{n}(Z_n-3) \rightsquigarrow N(0,24)$$

If y_τ comes from the second model in (9) (and assuming $\nu>4$) then

$$\frac{1}{n}\sum(y_i-\bar{y})^4 \to 3\phi_T^2\frac{\nu^2}{(\nu-2)(\nu-4)} \quad (a.s.)$$

and

$$\left(\frac{1}{n}\sum(y_i-\bar{y})^2\right) \to \phi_T\frac{\nu}{\nu-2} \quad (a.s.)$$

by explicit calculation of the moments of the t-distribution [1] and thus

$$Z_n \to 3.\frac{(\nu-2)}{(\nu-4)} \quad (a.s.).$$

As y_τ comes from second model of (9), ν is finite and as $n\to\infty$

$$\sqrt{n}(Z_n-3)$$

will diverge. □

References

[1] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*, 9th ed. Dover, New York.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory.* Akademiai Kiado, Budapest, Hungary, pp. 267–281.

[3] Aston, J. A. D. and Koopman, S. J. (2006). A non-Gaussian generalisation of the airline model for robust seasonal adjustment. *Journal of Forecasting* **25** 325–349.

[4] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd ed. Holden-Day, San Francisco, CA.

[5] Bruce, A. G. and Jurke, S. R. (1996). Non-Gaussian seasonal adjustment: X-12-ARIMA versus robust structural models. *Journal of Forecasting* **15** 305–328.

[6] Findley, D. F., Martin, D. E. K., and Wills, K. C. (2002). Generalizations of the Box–Jenkins airline model. In *Proceedings of the Business and Economics Section.* American Statistical Association [CD-ROM].

[7] Gómez, V. and Maravall, A. (1996). Programs TRAMO and SEATS: Instructions for the user. Working Paper 9628, Servicio de Estudios, Banco de España.

[8] Gómez, V. and Maravall, A. (2001). Automatic modeling methods for univariate series. In *A Course in Time Series*, D. Peña, G. C. Tiao, and R. S. Tsay, Eds. J. Wiley and Sons, New York, NY. Chapter 7.

[9] Hillmer, S. C. and Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* **77** 63–70.

[10] Kitagawa, G. (1989). Non-Gaussian seasonal adjustment. *Computers and Mathematics with Applications* **18** 503–14.

[11] Monsell, B. C., Aston, J. A. D., and Koopman, S. J. (2003). Towards X-13? In *Proceedings of the Business and Economics Section*. American Statistical Association [CD-ROM].

[12] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, UK.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 149–164

DOI: 10.1214/074921706000001012

Fractional constant elasticity of variance model*

Ngai Hang Chan[1] and Chi Tim Ng

The Chinese University of Hong Kong

Abstract: This paper develops a European option pricing formula for fractional market models. Although there exist option pricing results for a fractional Black-Scholes model, they are established without accounting for stochastic volatility. In this paper, a fractional version of the Constant Elasticity of Variance (CEV) model is developed. European option pricing formula similar to that of the classical CEV model is obtained and a volatility skew pattern is revealed.

1. Introduction

Black [3, 4] developed the so-called Constant Elasticity of Variance model for stock price processes that exhibit stochastic volatility. The CEV model is expressed in terms of a stochastic diffusion process with respect to a standard Brownian motion

$$dX_t = \mu X_t\,dt + \sigma X_t^{\beta/2}\,dB_t, \tag{1.1}$$

where $0 \le \beta \le 2$ is a constant. If $\beta = 2$, such a model degenerates to a geometric Brownian motion. This model is characterized by the dependence of the volatility rate, i.e., $\sigma X^{\beta/2}$ on the stock price. When the stock price increases, the instantaneous volatility rate decreases. This seems reasonable because the higher the stock price, the higher the equity market value, and thus the lower the proportion of liability, which results in a decrease in the risk of ruin. The volatility rate or the risk measure is thus decreased. Making use of methods proposed in an earlier literature [6], Cox [5] studied the CEV models and gave an option pricing formula that involves a noncentral χ^2 distribution function.

The classical CEV model (1.1) does not account for long-memory behavior, however. There are some evidences showing that the financial market exhibits long-memory structures (see [7, 23]). To encompass both long-memory and stochastic volatility, a possible model is to replace the Brownian motion in the stochastic diffusion equation by a fractional Brownian motion that exhibits a long-memory dependence structure (see [2, 13, 30]).

Though fractional Brownian motion can be used to model long-memory, as pointed out by Rogers in [29], the fractional Brownian motion is not a semimartingale and the stochastic integral with respect to it is not well-defined in the classical Itô's sense. A theory different from the Itô's one should be used to handle the fractional situation. One approach is white noise calculus (see [18, 22, 28]),

*Research supported in part by HKSAR RGC grants 4043/02P and 400305.
[1]The Chinese University of Hing Kong, e-mail: nhchan@sta.cuhk.edu.hk
AMS 2000 subject classifications: primary 91B28, 91B70; secondary 60H15, 60H40.
Keywords and phrases: fractional Black-Scholes model, fractional Brownian motion, fractional constant elasticity of volatility model, fractional Itô's lemma, volatility skew, white noise, Wick calculus.

which was used in [14, 19] to construct stochastic integral with respect to the fractional Brownian motion. With the white noise approach, an extension to the Black-Scholes' stochastic differential equation is proposed to cope with long-memory phenomena (see [26]).

In this paper, the white noise calculus approach is adopted to construct a fractional CEV model and to derive the option pricing formula for European call option. Basic concepts in white noise calculus are briefly introduced in Section 2. The fractional Itô's lemma, which is fundamental to option pricing theory, is presented in Section 3. Section 4 explains under what circumstances is the Itô's lemma applicable. In Section 5, the fractional option pricing theory is introduced and the concept of self-financing strategy, which is different from the traditional definition adopted by, e.g., Delbaen [10, 11], will further be discussed. Finally, the pricing formula for fractional CEV model is given in Sections 6.

2. White noise calculus and stochastic integration

In this section, the concept of stochastic integration with respect to fractional Brownian motion is introduced briefly. Important concepts are defined based on the white noise theory originated from [17], who considered the sample path of a Brownian motion as a functional. Throughout this paper, notations used in [1, 14, 18, 22] are adopted.

Let $S(R)$ be the Schwarz space. Take the dual $\Omega = S'(R)$, equipped with the weak star topology, as the underlying sample space, i.e., $\omega \in \Omega$ is a functional that maps a rapidly decreasing function $f(\cdot) \in S(R)$ to a real number. Also, let $B(\Omega)$ be the σ-algebra generated by the weak star topology. Then according to Bochner-Minlos Theorem (see Appendix A of [18]), there exists a unique probability measure μ on $B(\Omega)$, such that for any given $f \in S(R)$, the characteristic function of the random variable $\omega \to \omega(f)$ is given by

$$\int_\Omega e^{i\omega(f)}\, d\mu(\omega) = e^{-\frac{1}{2}||f||^2},$$

where

$$||f||^2 = \int_R f^2(t)\, dt.$$

Let L^2 be the space of real-valued functions with finite square norm $||\cdot||$, we have the triple $S(R) \subset L^2 \subset S'(R)$. For any $f \in L^2$, we can always choose a sequence of $f_n \in S(R)$ so that $f_n \to f$ in L^2, and $\omega(f)$ is defined as the $\lim_{n\to\infty} \omega(f_n)$ in $L^2(\mu)$.

Consider the indicator function

$$1_{(0,a)}(s) = \begin{cases} 1, \text{ if } 0 \le s < a, \\ -1, \text{ if } a \le s < 0, \\ 0, \text{ otherwise.} \end{cases}$$

It can be verified that for any two real numbers a and b, the random variables $\omega(1_{(0,a)}(\cdot))$ and $\omega(1_{(0,b)}(\cdot))$ are jointly normal, mean zeros, and with covariance $\min(a, b)$. Define $\tilde{B}(t)$ as $\omega(1_{(0,t)}(\cdot))$, we can always find a continuous version of $\tilde{B}(t)$, denoted by B, which is the standard Brownian motion. Roughly speaking, the probability space $(\Omega, B(\Omega), \mu)$ can intuitively be considered as a space consisting of all sample paths of a Brownian motion.

Following the approach of [14], we give the definition of fractional Brownian motion with Hurst parameter $\frac{1}{2} \leq H < 1$ in terms of white noise setting by using the fundamental operator M_H, defined on the space L^2, by

$$M_H f(x) = \begin{cases} f(x), & H = \frac{1}{2}, \\ c_H \int_R f(t)|t-x|^{H-\frac{3}{2}}\, dt, & H < \frac{1}{2} < 1, \end{cases}$$

where c_H is a constant depending on the Hurst parameter H via

$$c_H = (2\Gamma(H-\frac{1}{2})\cos(\frac{\pi}{2}(H-\frac{1}{2})))^{-1}[\Gamma(2H+1)\sin(\pi H)]^{\frac{1}{2}}.$$

Then, $\omega(M_H 1_{(0,a)}(\cdot))$ and $\omega(M_H 1_{(0,b)}(\cdot))$ are jointly normal with covariance

$$\frac{1}{2}\{|a|^{2H} + |b|^{2H} - |a-b|^{2H}\}.$$

Again, we can find a continuous version of $\omega(M_H 1_{(0,t)}(\cdot))$ denoted by $B^H(t)$, which is the fractional Brownian motion.

We have the following Wiener-Itô Chaos Decomposition Theorem for a square integrable random variable on $S'(R)$ (see Theorem 2.2.4 of [18]).

Theorem 2.1. *If $F \in L^2(\Omega, B(\Omega), \mu)$, then $F(\omega)$ has a unique representation*

$$F(\omega) = \sum_{\alpha} c_\alpha H_\alpha(\omega),$$

where α is any finite integers sequence $(\alpha_1, \alpha_2, \ldots, \alpha_n)$, c_α are real coefficients and $H_\alpha(\omega) = h_{\alpha_1}(\omega(e_1))h_{\alpha_2}(\omega(e_2))\cdots h_{\alpha_n}(\omega(e_n))$ $h_n(x)$ are Hermite polynomials and e_n is an orthonormal set in $S(R)$ which is defined as

$$e_i(t) = (\sqrt{\pi}2^{i-1}(i-1)!)^{-1/2} h_{i-1}(t) e^{-t^2/2}.$$

Furthermore, the L^2 norm of the functional $F(\omega)$ is given by

$$\sum_{\alpha} \alpha! c_\alpha^2,$$

with $\alpha! = \alpha_1!\alpha_2!\cdots\alpha_n!$.

Remark 2.1. (see [18, 27]) The basis $\{H_\alpha(\omega) : \alpha\}$ is orthogonal with respect to the inner product $E(XY)$ in $L^2(\Omega, B(\Omega), \mu)$. The variance of $H_\alpha(\omega)$ is $\alpha!$. $H_0(\omega)$ is taken as the constant 1. For $\alpha \neq 0$, the expectation of $H_\alpha(\omega)$ is

$$E(H_\alpha(\omega)H_0(\omega)) = 0.$$

As a result, the term c_0 is the expectation of the functional $F(\omega)$.

Consider the functional $B_t^H = \omega(M_H 1_{[0,t]}(.))$, where t is a given constant. Using the dual property (see [14]) of the M_H operator: for all rapidly decreasing functions f and g, we have

$$(f, M_H g) \equiv \int_R f(t) M_H g(t)\, dt = \int_R g(t) M_H f(t)\, dt \equiv (M_H f, g).$$

The function $M_H 1_{[0,t]}(.)$ can be rewritten in the Fourier expansion form:

$$\begin{aligned} M_H 1_{[0,t]}(s) &= \sum_{i=0}^{\infty} (M_H 1_{[0,t]}(\cdot), e_i(\cdot)) e_i(s) \\ &= \sum_{i=0}^{\infty} (1_{[0,t]}(\cdot), M_H e_i(\cdot)) e_i(s) \\ &= \sum_{i=0}^{\infty} \left\{ \int_0^t M_H e_i(u) du \right\} e_i(s). \end{aligned}$$

Since ω is linear, the functional can be written as

$$\begin{aligned} \omega(M_H 1_{[0,t]}(\cdot)) &= \sum_{i=0}^{\infty} (1_{[0,t]}(\cdot), e_i(\cdot)) \omega(e_i) \\ &= \sum_{i=0}^{\infty} \left\{ \int_0^t M_H e_i(u) du \right\} \omega(e_i). \\ &= \sum_{\alpha} c_\alpha H_\alpha(\omega). \end{aligned}$$

In this example, when $\alpha = \epsilon(i) = \{0..., 1, 0, ...\}$, i.e., one at position i, $c_\alpha = \int_0^t M_H e_i(u) du$, and $c_\alpha = 0$ otherwise. It is tempting to write

$$\frac{d}{dt} B_t^H = \sum_{i=0}^{\infty} M_H e_i(t) \omega(e_i),$$

which is illegitimate in the traditional sense as the Brownian motion or the fractional Brownian motion is nowhere differentiable. With the chaos expansion form, differentiation and integration with respect to time t can be defined, but they may not always be square integrable. Such type of operation is called integration or differentiation in $(S)^*$ (see [18]).

Definition 2.1. Let (S) be a subset of $L^2(\Omega, B(\Omega), \mu)$ consisting of functionals with Wiener-Itô Chaos decomposition such that

$$\sum_{\alpha} \left\{ c_\alpha^2 \alpha! \prod_{j \in N} (2j)^{k\alpha_j} \right\} < \infty$$

for all $k < \infty$ and that $(S)^*$ consists of all expansions, not necessarily belonging to $L^2(\Omega, B(\Omega), \mu)$, such that

$$\sum_{\alpha} \left\{ c_\alpha^2 \alpha! \prod_{j \in N} (2j)^{-q\alpha_j} \right\} < \infty$$

for some $q < \infty$, then, the spaces (S) and $(S)^*$ are called the Hida test function space and the Hida distribution space respectively.

The derivative of the fractional Brownian motion, or the white noise is defined by

$$W^H(t) = \sum_{i=0}^{\infty} M_H e_i(t) \omega(e_i).$$

It can be shown that the sum $W^H(t) \in (S)^*$ (see [14, 31]). The importance of Hida test function space and distribution space is their closedness of Wick multiplication. Wick product is an operator acting on two functionals $F(\omega)$ and $G(\omega)$.

Definition 2.2. The Wick's product for two functionals having Wiener-Itô Chaos Decomposition

$$F(\omega) = \sum_{\alpha} c_\alpha H_\alpha(\omega)$$

and

$$G(\omega) = \sum_{\beta} b_\beta H_\beta(\omega)$$

is defined as

$$F(\omega) \diamond G(\omega) = \sum_{\alpha,\beta} c_\alpha b_\beta H_{\alpha+\beta}(\omega).$$

Addition of indexes refers to pairwise addition.

The closeness of Wick's product is shown in the following theorem (see Corollary 2.2 and Remark 2.8 of [31]).

Theorem 2.2. *Wick multiplication is closed in (S) and* $(S)^*$.

It is reasonable to define the stochastic integration of a functional $Z_t(\omega)$ with respect to the fractional Brownian motion as the integration with respect to time t of the Wick's product between $Z_t(\omega)$ and $W_t^H(\omega)$. Under the Wiener Chaos decomposition framework, if the decomposition exists, the functional $Z_t(\omega) \diamond W_t^H(\omega)$ can be written as

$$\sum_{\alpha} c_\alpha(t) H_\alpha(\omega).$$

It is natural to think that the integration is

$$\sum_{\alpha} \left\{ \int_0^t c_\alpha(s) ds \right\} H_\alpha(\omega),$$

by assuming that summation and integration are interchangeable. If the integration with respect to time is a path-wise classical Riemann integral, it is not clear that summation and integration are interchangeable. Such difficulties can be finessed by introducing new definitions for integration with respect to time and integration with respect to the fractional Brownian motion as follows (see Definitions 2.3 and 2.4 respectively).

Definition 2.3. (a) Elements in $(S)^*$ as an operator: Let $F(\omega) = \sum_\alpha c_\alpha H_\alpha(\omega) \in (S)^*$ and $f(\omega) = \sum_\alpha b_\alpha H_\alpha(\omega) \in (S)$, then F can be regarded as an operation on f

$$\langle F, f \rangle = \sum_{\alpha} b_\alpha c_\alpha \alpha!.$$

(b) Time integration: If $F_t(\omega) = \sum_\alpha c_\alpha(t) H_\alpha(\omega)$ are elements in $(S)^*$ for all positive real number t, and that $\langle F_t, f \rangle$ are integrable with respect to t for all $f \in (S)$, then the integral $\int_R F_t(\omega) dt$ is defined as the unique element in $(S)^*$, $I(\omega)$ such that

$$\langle I(\omega), f \rangle = \int_R \langle F_t(\omega), f \rangle \, dt$$

for all $f \in (S)$.

Remark 2.2. It can be shown that the quantity $\langle F, f\rangle$ in part (a) exists and is finite under the condition given in the definition. It can be regarded as the expectation of the product between $F(\omega)$ and $f(\omega)$ when $F(\omega) \in L^2(\Omega, B(\Omega), \mu)$ and $f(\omega) \in (S)$.

Definition 2.4. Let $Z(t) = \sum_\alpha c_\alpha(t)H_\alpha(\omega) \in (S)^*$ for any given t, then, the Wick's integral of $Z(t)$ is defined as

$$\int_R Z(t) \diamond dB^H(t) = \int_R \{Z(t) \diamond W^H(t)\}\, dt,$$

when $Z(t) \diamond W^H(t)$ is integrable with respect to time t in the sense as in Definition 2.3.

The following theorem asserts that integration and summation are interchangeable, see Lemma 2.5.6 of [18].

Theorem 2.3. *Let $Z(t) : R \to (S)^*$, with Wiener Chaos decomposition*

$$\sum_\alpha c_\alpha(t)H_\alpha(\omega)$$

such that

$$\sum_\alpha \alpha!\{\int_R c_\alpha(t)dt\}^2 \prod_{j\in N}(2j)^{-q\alpha_j} < \infty$$

for some $q < \infty$, then $Z(t)$ is time-integrable, also, integration and summation are interchangeable, i.e.

$$\int_R Z(t)dt = \sum_\alpha \{\int_R c_\alpha(t)dt\}H_\alpha(\omega).$$

3. Fractional Itô's lemma

Several approaches of extending classical Itô's lemma to incorporate the fractional Brownian motion were discussed in the literature, e.g., [8, 9, 12]. The settings in these papers are different and various conditions are required to ensure that the stochastic integrals appear in the fractional Itô lemma exist. Bender in [1] provided a simpler version of fractional Itô's lemma based on the white noise setting introduced in the preceding section. Here, we restate Bender's theorem in Theorem 3.1 and give a generalized result in Theorem 3.2.

Theorem 3.1. *Consider the stochastic process*

$$Y_t = \int_0^t h(t)\, dB_t^H \equiv \omega[M_H(h(\cdot)1_{[0,t)}(\cdot))],$$

where $h(t)$ is a continuous function in $[0, T]$ and $H > \frac{1}{2}$. Let $g(t, y)$ be a two-dimensional function differentiable with respect to t and is twice differentiable with respect to y. Also, there exists constants $C_1 \geq 0$ and $\lambda_1 < (2T^H \sup_{s\in[0,T]} h(s))^{-2}$ so that

$$\max\{|g|, |g_t|, |g_y|, |g_{yy}|\} \leq C_1 e^{\lambda_1 y^2}.$$

Let $k(t) = H(2H-1)h(t)\int_0^T h(s)|s-t|^{2H-2}\,ds$. *Then, we have the following fractional Itô's lemma:*

$$g(T,Y_T) = g(0,0) + \int_0^T \frac{\partial}{\partial t}g(t,Y_t)\,dt + \int_0^T h(t)\frac{\partial}{\partial y}g(t,Y_t)\diamond dB_t^H + \int_0^T k(t)\frac{\partial^2}{\partial y^2}g(t,Y_t)\,dt.$$

Remark 3.1. The integrals above are well defined as the condition that the integrands and integrals both belong to (L^2) is ensured by the given assumption. Since $(L^2)\subset(S)^*$, the integrals are defined as in Section 2.

This theorem gives the differential form of $g(t,Y_t)$ when the underlying stochastic process Y_t is an integrals of a deterministic function h_t. The following generalization takes the underlying stochastic process to be $X_t = g(t,Y_t)$ and gives the differential form for $P(t,X_t)$, where $P(t,x)$ is a two-dimensional real-valued function. Before introducing our results, let us illustrate some ideas through an example (see [19]).

Consider the two-dimensional function,

$$g(t,y) = \exp(\mu t - \frac{1}{2}\sigma^2 t^{2H} + \sigma y),$$

where μ and σ are two positive constants and the underlying stochastic process is

$$Y_t = \int_0^t dB_s^H,$$

i.e., $h(t)=1$. Clearly, for any given value of T, the functions g, g_t, g_y and g_{yy} are all continuous in the closed interval $[0,T]$ and hence, the conditions in the theorem are fulfilled. Applying the lemma, we have

$$k(t) = H(2H-1)\int_0^t |t-s|^{2H-2}\,ds = Ht^{2H-1},$$

and

$$\begin{aligned} dX_t &= (\mu - \sigma^2 Ht^{2H-1})g(t,Y_t)\,dt + \sigma g(t,Y_t)\diamond dB_t^H \\ &\quad + k(t)\sigma^2 g(t,X_t)\,dt \\ &= \mu g(t,Y_t)\,dt + \sigma g(t,Y_t)\diamond dB_t^H \\ &= \mu X_t\,dt + \sigma X_t \diamond dB_t^H. \end{aligned}$$

The next question is whether there exists an Itô's lemma that further expresses $P(t,X_t)$ in terms of integrals involving μX and σX, but not Y and g explicitly. It is reasonable to expect that

$$dP(t,X_t) = P_t(t,X_t)\,dt + \mu Y_t P_x(t,X_t)\,dt + \sigma Y_t P_x(t,Y_t)\diamond dB_t^H + \sigma^2 Ht^{2H-1}X^2 P_{xx}(t,X_t)\,dt.$$

This result can be verified by the following theorem.

Theorem 3.2. *Using the same notations and assumptions in Theorem 3.1, further assume that the differential of the stochastic process* $X_t = g(t,Y_t)$ *can be, according to Theorem 3.1, written as*

$$dX_t = \mu(t,X_t)\,dt + \sigma(t,X_t)\diamond dB_t^H.$$

Also, there exists constants $C_2 \geq 0$ *and* $\lambda_2 < (2T^H \sup_{s\in[0,T]} h(s))^{-2}$ *so that the composite function* $P \circ g \equiv P(t, g(t, y))$ *satisfies*

$$\max\{|(P \circ g)|, |(P \circ g)_t|, |(P \circ g)_y|, |(P \circ g)_{yy}|\} \leq C_2 e^{\lambda_2 y^2}.$$

Let $C(t) = k(t)h^{-2}(t)$. *Then the fractional Itô's lemma is given by*

$$P(T, X_T) = P(0, X_0) + \int_0^T \frac{\partial}{\partial t} P(t, X_t)\, dt + \int_0^T \mu(t, Y_t) \frac{\partial}{\partial x} P(t, X_t)\, dt + \int_0^T \sigma(t, Y_t) \frac{\partial}{\partial x} P(t, X_t) \diamond dB_t^H + \int_0^T C(t)\sigma^2(t, Y_t) \frac{\partial^2}{\partial x^2} P(t, X_t)\, dt.$$

Proof. This theorem can be verified by applying Theorem 3.1 to $f(t, g(t, Y_t))$. □

To relate this result with previous works on fractional stochastic calculus and option pricing theory, such as [12, 26], consider the last term in the right-hand side of the Itô's formula. This second order correction term can be considered as the product of three quantities: $P_{xx}(t, X)$, $\sigma(t, X)$ and $C(t)\sigma(t, X)$. Comparing with the result of [12], the quantity $C(t)\sigma(t, X)$ corresponds to $D_t^\phi X_t$. This quantity is known as the Malliavin derivative of X_t. For details of Malliavin calculus, see [20, 24, 27]. Under our assumptions that X_t has the form of $g(t, Y_t)$, this quantity depends only on the current time t and the value of X at time t. But this needs not be the situation for a general X_t governed by a fractional stochastic differential equation. In general this quantity may depend on the entire path of X_t, not only on the value at one point. This may introduce further complications when working with the differential.

4. The fractional CEV model

Here we construct a fractional version of the Constant Elasticity of Variance model by means of the fractional Itô's lemma (Theorem 3.1). As discussed in the introduction, constant elasticity is characterized by the the volatility term $\sigma X_t^{\beta/2}$ in a stochastic diffusion equation. In order to handle long-memory, we replace the Wiener process by a fractional Brownian motion. The fractional diffusion equation is then defined as

$$dX_t = \mu(t, X_t)\, dt + \sigma X_t^{\beta/2} \diamond dB_t^H,$$

where $0 \leq \beta \leq 2$. If $H = \frac{1}{2}$ and $\mu(t, X_t) \equiv \mu X_t$, this is the classical CEV model. In this situation, the Wick integral is equivalent to the Itô's integral (see [18]) and hence, the classical Itô's lemma can be applied to any stochastic process of the form $Y_t = P(t, X_t)$. When long-memory is considered, the Itô's lemma will involve the Malliavin derivative, which is in general path dependent and difficult to handle. When the integration is defined in the white noise sense (Section 2), the integrand is assumed to belong to $(S)^*$ at every time t, and the integral is a random variable in $(S)^*$. The elements in $(S)^*$, which are merely formal expansions, may not correspond to real values for each ω and the term $P(t, X_t)$ may not be well-defined in general. In order to overcome such difficulties, we need to choose a suitable $\mu(t, X_t)$.

Assume that X_t can be written explicitly in terms of time t and a stochastic integral process $Y_t = \int_0^t h(s) dB_s^H$, i.e. $X_t = g(t, Y_t)$. From Theorem 3.1, the differential of X_t can be decomposed into two parts, the drift term and the volatility

term. In order to keep the elasticity constant, the function $g(t,y)$ must be chosen as

$$g(t,y) = \left\{\sigma(1-\frac{\beta}{2})[h(t)]^{-1}y + f(t)\right\}^{\frac{2}{2-\beta}},$$

where $f(t)$ is an arbitrarily chosen function of t. This $g(t,y)$ is in fact the solution to the differential equation

$$h(t)\frac{\partial}{\partial y}g(t,y) = \sigma\{g(t,y)\}^{\beta/2}.$$

The next question is how to choose $f(t)$. The answer is given by the following theorem.

Theorem 4.1. *Assume that $h(t)$ is a strictly positive function and $g(t,y)$ is the solution to the differential equation*

$$h(t)\frac{\partial}{\partial y}g(t,y) = \sigma(t,g(t,y)). \tag{4.1}$$

The general solution to this equation involves an arbitrary function $f(t)$.

Let $\eta(t)$ and $\varphi(t)$ be two functions determined by the integral equations,

$$h(t) = e^{-\int_0^t \eta(s)ds}$$

and

$$f(t) = [h(t)]^{-1}\left[a_0 + \int_0^t h(s)\varphi(s)\,ds\right],$$

where a_0 is a constant so that $g(0,0) = X_0$, then,

$$X_t = g(t, Y_t)$$

yields the volatility $\sigma(t,X_t)$ in the fractional Itô's lemma and the drift term is given by a two dimensional function

$$\mu(t,x) = \sigma(t,x)\left[\varphi(t) + C(t)\,\frac{\partial\sigma}{\partial x} + \int_0^x \frac{\eta(t)}{\sigma(t,x)}\,dx + \int_0^x \frac{1}{\sigma^2(t,x)}\,\frac{\partial\sigma}{\partial t}\,dx\right].$$

Proof. Let

$$a(t,x) = \int \frac{dx}{\sigma(t,x)},$$

then $g(t,y)$ given by

$$a(t,g(t,y)) = [h(t)]^{-1}y + f(t)$$

satisfies Equation (4.1). After some manipulations, we have

$$\begin{aligned}&\frac{\partial}{\partial t}g(t,y) + k(t)\frac{\partial^2}{\partial y^2}g(t,y)\\ &\quad = \sigma(t,g)\left[\varphi(t) + C(t)\frac{\partial\sigma}{\partial x}(t,g) + \eta(t)a(t,g) - \frac{\partial a}{\partial t}\right],\end{aligned}$$

which does not involve y explicitly. The proof is completed by comparing this with fractional Itô's lemma (Theorem 3.1). □

Remark 4.1. Note that the function $C(t)$ depends on the choice of $\eta(t)$ and by definition (in Theorem 3.2) must be strictly positive.

In the CEV case, when the drift term has the form of

$$\mu(t,x) = \frac{\eta(t)}{1-\beta/2}x + \sigma\varphi(t)x^{\beta/2} + \frac{\sigma^2\beta}{2}x^{\beta-1}C(t), \tag{4.2}$$

the solution to the stochastic diffusion equation

$$dX_t = \mu(t,X_t)\,dt + \sigma X_t^{\beta/2} \diamond dB_t^H$$

is well defined and is given by Theorem 4.1. Now choose suitable time dependent functions $\eta(t)$ and $\varphi(t)$. Consider the three quantities of $\mu(t,X_t)$ in (4.2). The last one is strictly positive and cannot be eliminated. When we choose $\eta(t) \equiv (1-\frac{\beta}{2})\mu$ and $\varphi(t) \equiv 0$, the second term vanishes and the first term becomes μX_t. The drift term becomes

$$\mu(t,X_t) = \mu X_t + \frac{\sigma^2\beta}{2}C(t)X_t^{\beta-1}.$$

The CEV stochastic differential equation is thus

$$X_t = X_0 + \int_0^t \mu X_s\,ds + \left\{\int_0^t \frac{\sigma^2\beta}{2}C(s)X_s^{\beta-1}\,ds + \int_0^t \sigma X_s^{\beta/2} \diamond dB_s^H\right\}.$$

5. Fractional option pricing theory

By using the generalized Itô's lemma (Theorem 3.2), the differential of $P(t,X_t)$ can be decomposed into two terms, the drift one and the volatility one and both terms involve only current time t and X_t, i.e., they are path-independent. The foundation of the Black-Scholes' option pricing theory is constructing a self-financing strategy, which makes use of stocks and bonds to hedge an option. The definition for self-financing strategy in continuous-time model depends on how the stochastic integrals are defined. As the fractional stochastic integrals are defined in a different manner, a new definition for self-financing strategy is required. One approach is adopted by [14, 19], which defines self-financing strategy under the geometric fractional Brownian motion. Here, this approach is extended to a more general situation.

Let X_t be the stock price process and Π_t be the bond value and they are governed by

$$\begin{aligned} X_t &= \int_0^t \mu(s,X)\,ds + \int_0^t \sigma(s,X) \diamond dB_s^H, \\ \Pi_t &= \int_0^t r\Pi\,ds. \end{aligned} \tag{5.1}$$

Definition 5.1. (see Section 5 of [14]) A trading strategy consists of a quantity (u_t, v_t) of bonds and stocks is called self-financing if the infinitesimal change in the portfolio value at time t is given by

$$\begin{aligned} dZ_t &= d(u_t\Pi_t + v_tX_t) \\ &= r\Pi_t u_t\,dt + \mu(t,X_t)v_t\,dt + [\sigma(t,X_t)v_t] \diamond dB_t^H + d\Delta, \end{aligned}$$

where $d\Delta$ is an infinitesimal dividend payment term.

Below, using the above definition, a fractional Black-Scholes equation is derived.

Theorem 5.1. *Suppose that the market consists of two securities, a risk-free bond and a stock. Here, the stock provides dividend continuously with rate δ. Assume that the stock price process $X_t = g(t, Y_t)$ is defined as in Section 3 which also satisfies the equations (5.1). Then the price of a derivative on the stock price with a bounded payoff $f(X(T))$ is given by $P(t, X)$, where $P(t, X)$ solves the PDE:*

$$\frac{\partial P}{\partial t} + \sigma^2(t, X)C(t)\frac{\partial^2 P}{\partial X^2} + (r - \delta)X\frac{\partial P}{\partial X} - rP = 0, \tag{5.2}$$

with boundary condition $P(T, X) = f(X)$ given that $P \circ g$ satisfies the conditions in Theorem 3.2.

Proof. A proof parallel to the fractional Black-Scholes equation in [26] is given here. Consider a solution $P(t, X)$ to equation (5.2). Applying fractional Itô's Lemma Theorem 3.2,

$$dP(t, X_t) = [\frac{\partial P}{\partial t} + \frac{\partial P}{\partial X}\mu(t, X) + \frac{\partial^2 P}{\partial X^2}\sigma^2(t, X)C(t)]\, dt + \frac{\partial P}{\partial X}\sigma(t, X) \diamond dB_t^H.$$

Form a trading strategy by dynamically adjusting a portfolio consisting a varying quantity $v(t)$ of stocks and $u(t)$ of bonds. By choosing

$$\begin{aligned} v(t) &= \frac{\partial P}{\partial X} \\ u(t) &= \frac{1}{\Pi_t}(P - X\frac{\partial P}{\partial X}), \end{aligned} \tag{5.3}$$

then the portfolio value at time t is P_t and

$$\begin{aligned} &ru(t)\Pi_t\, dt + v(t)\mu(t, X_t)\, dt + [v(t)\sigma(t, X_t)] \diamond dB_t^H + \delta v(t)X_t\, dt \\ &\quad = [rP - rX\frac{\partial P}{\partial t}]\, dt + \mu\frac{\partial P}{\partial X}\, dt + (\sigma\frac{\partial P}{\partial X}) \diamond dB_t^H + \frac{\partial P}{\partial X}\delta X\, dt \\ &\quad = [\frac{\partial P}{\partial t} - \frac{\partial P}{\partial X}\,\delta X\, dt + \frac{\partial^2 P}{\partial X^2}\sigma^2(t, X)C(t)]\, dt \\ &\qquad + \mu\frac{\partial P}{\partial X}\, dt + (\sigma\frac{\partial P}{\partial X}) \diamond dB_t^H + \frac{\partial P}{\partial X}\,\delta X\, dt \\ &\quad = [\frac{\partial P}{\partial t} + \frac{\partial P}{\partial X}\mu(t, X) + \frac{\partial^2 P}{\partial X^2}\sigma^2(t, X)C(t)]\, dt + \frac{\partial P}{\partial X}\sigma(t, X) \diamond dB_t^H \\ &\quad = dP(t, X_t) \\ &\quad = d\,(u_t\Pi_t + v_tX_t). \end{aligned}$$

By Definition 5.1, $(u(t), v(t))$ is a self-financing strategy. It can be shown that such strategy hedges the derivative. The portfolio value at time t is given by $u(t)\Pi_t + v(t)X_t$ and it is equal to $P(t, X_t)$. At time of maturity, the portfolio value is just $P(T, X_T)$. By assumption, the function $P(t, X)$ satisfies the boundary condition, so $P(T, X_T) = f(X_T)$. Therefore $(u(t), v(t))$ hedges the derivative and $P(t, X)$ is the option price. □

6. Pricing an European call option under CEV models

Putting $\sigma^2(t, X) \equiv \sigma^2 X^\beta$, the Black-Scholes PDE (Theorem 5.1) of the CEV model is now given by

$$\frac{\partial P}{\partial t} + \sigma^2 C(t)X^\beta\frac{\partial^2 P}{\partial X^2} + (r - \delta)X\frac{\partial P}{\partial X} - rP = 0.$$

Putting $Y = X^{2-\beta}$ (see [5]) and $P(t, X) = e^{rt}Q(t, Y)$, this equation becomes

$$\frac{\partial Q}{\partial t} + [bY + cC(t)]\frac{\partial Q}{\partial Y} + aC(t)Y\frac{\partial^2 Q}{\partial Y^2} = 0,$$

where $a = \sigma^2(2-\beta)^2$, $b = (r-\delta)(2-\beta)$ and $c = \sigma^2(2-\beta)(1-\beta)$. The boundary condition is

$$Q(T, Y) = e^{-rT}\max(Y^{\frac{1}{2-\beta}}, 0).$$

The approach of Cox and Ross [6] that made use of Feller's result ([15, 16]) can be adopted. First the solution for this equation at (t_0, Y_{t_0}) is the expectation of $Q(T, Y_T)$ under the SDE

$$dY = [bY + cC(t)]\,dt + \sqrt{2aC(t)Y}\,dB_t, \tag{6.1}$$

with $Y(t_0) = Y_{t_0}$ (see [25]). To solve this SDE, we follow Feller's arguments. First, a useful result of Kolmogorov is stated below (see Equation (167) of [21]).

Theorem 6.1. *The probability density function of a diffusion process X_t driven by standard Brownian motion*

$$dX_t = \mu(t, X_t)\,dt + \sigma(t, X_t)dB_t$$

is given by the PDE

$$u_t = [\frac{1}{2}\sigma^2(t, X)u(t, X)]_{XX} - [\mu(t, X)u(t, X)]_X.$$

In our case, because of (6.1), the Kolmogorov's equation becomes

$$u_t = [aC(t)Yu(t, Y)]_{YY} - [(bY + cC(t))u(t, Y)]_Y.$$

The European call option pricing formula can be obtained by solving the above PDE.

Theorem 6.2. *Under the fractional CEV model introduced in Section 4, the price of an European call option with strike price K, mature at T at current time t_0 is given by*

$$\begin{aligned} P(t_0, X_0) = {} & e^{-\delta(T-t_0)}X_0\sum_{r=0}^{\infty}\frac{1}{r!}e^{-\frac{x}{a\gamma_T}}(\frac{x}{a\gamma_T})^r G(r+1+\frac{1}{2-\beta}, \frac{K^{2-\beta}}{ae^{bT}\gamma_T}) \\ & - Ke^{-r(T-t_0)}\sum_{r=0}^{\infty}\frac{1}{\Gamma(r+1-\frac{1}{2-\beta})}e^{-\frac{x}{a\gamma_T}}(\frac{x}{a\gamma_T})^{r+\frac{1}{2-\beta}}G(r+1, \frac{K^{2-\beta}}{ae^{bT}\gamma_T}), \end{aligned}$$

where $x = e^{-bt_0}Y(t_0)$,

$$\gamma_t \equiv \int_{t_0}^{t} e^{-b\tau}C(\tau)\,d\tau$$

and

$$G(\alpha, \nu) \equiv \frac{1}{\Gamma(\alpha)}\int_{\nu}^{\infty} e^{-\tau}\tau^{\alpha-1}\,d\tau.$$

Proof. Assume that the Laplace Transform of $u(t, Y)$ with respect to Y exists and equals to $\omega(t, s)$. Since the value of Y at time t_0 is given, Y_{t_0} is deterministic and

thus $u(t_0, Y) = \delta(Y - Y_{t_0})$, the Dirac function. Also, $\omega(t_0, s) = e^{-sY_0} = \pi(s)$ and the equation becomes the boundary value problem

$$\omega_t + s(aC(t)s - b)\omega_s = -cC(t)s\omega + \psi(t) \tag{6.2}$$

$$\omega(t_0, s) = e^{-sY_0}. \tag{6.3}$$

In equation (6.2), $\psi(\cdot)$ is called the flux of u at the origin (see [16]), which is to be determined later. Now, we find the characteristic curve of the first order PDE (6.2). The characteristic curve is given by

$$\frac{ds}{dt} = aC(t)s^2 - bs.$$

The solution to this equation is

$$s = e^{-bt}[C_1 - a\gamma_t]^{-1}.$$

On this curve, $\omega(t, s(t))$ satisfies

$$\frac{d}{dt}\omega(t, s(t)) = \psi(t) - c\,C(t)s(t)\omega(t, s(t)).$$

Solving, we have

$$\omega(t, s(t)) = [C_1 - a\gamma_t]^{c/a}[C_2 + \int_{t_0}^{t} \psi(\tau)|C_1 - a\gamma_\tau|^{-c/a}\,d\tau].$$

For any given point (t_1, s_1), the characteristic curve with

$$C_1 = a\gamma_{t_1} + \frac{1}{s_1 e^{bt_1}} = C(t_1, s_1)$$

passes through (t_1, s_1). Also, this curve passes through the point $(t_0, C_1^{-1}e^{-bt_0})$. This yields the value of C_2,

$$C_2 = [C_1(t_1, s_1)]^{-c/a}\omega(t_0, e^{-bt_0}C_1^{-1}(t_1, s_1)).$$

The Laplace transform of $u(t, Y)$ at point (t_1, s_1) is thus given by

$$\begin{aligned}\omega(t_1, s_1) &= (s_1 e^{bt_1})^{-c/a}[(C_1(t_1, s_1))^{-c/a}\omega(t_0, e^{-bt_0}C_1^{-1}(t_1, s_1)) \\ &\quad + \int_{t_0}^{t_1} \psi(\tau)|a(\gamma_{t_1} - \gamma_\tau) + \frac{1}{s_1 e^{b\tau}}|^{-c/a}\,d\tau] \\ &= [s_1 a e^{bt_1}\gamma_{t_1} + 1]^{-c/a}\pi(\frac{s_1 e^{b(t_1 - t_0)}}{s_1 a e^{bt_1}\gamma_{t_1} + 1}) \\ &\quad + \int_{t_0}^{t_1} [s_1 a e^{bt_1}(\gamma_{t_1} - \gamma_\tau) + 1]^{-c/a}\psi(\tau)\,d\tau.\end{aligned} \tag{6.4}$$

Following the argument of [16], when $u(t, 0)$ is finite and $c \le 0$ or $0 < c < a$,

$$\lim_{s\to}(sae^{bt}\gamma_t + 1)\omega(t, s) = 0,$$

then $\psi(t)$ is given by the integral equation

$$\pi(\frac{e^{-bt_0}}{a\gamma_t}) + \int_{t_0}^{t} \psi(\tau)(\frac{\gamma_t}{\gamma_t - \gamma_\tau})^{c/a}\,d\tau = 0.$$

To solve this equation, applying the substitutions $z = \frac{1}{\gamma_t}$ and $\zeta = \frac{1}{\gamma_\tau}$,

$$\begin{aligned}\int_z^\infty g(\zeta)(\zeta - z)^{-c/a}\, d\zeta &= \pi(\frac{e^{-t_0}}{a\gamma_t}) = \pi(\frac{ze^{-t_0}}{a}) \\ &= \exp(-\frac{xz}{a}),\end{aligned} \tag{6.5}$$

where $g(\zeta) = -\psi(\tau)\zeta^{c/a}\frac{d\tau}{d\zeta}$.

The solution of (6.5) is

$$\begin{aligned} g(\zeta) &= \frac{1}{\Gamma(1-\frac{c}{a})}(\frac{x}{a})^{\frac{-c+a}{a}} \exp(-\frac{x}{a\gamma_\tau}), \\ \psi(\tau) &= g(\zeta)\zeta^{-c/a}\frac{d\zeta}{d\tau} \\ &= \frac{-1}{\Gamma(1-\frac{c}{a})}(\frac{1}{\gamma_\tau})(\frac{x}{a\gamma_\tau})^{\frac{-c+a}{a}} \exp(-\frac{x}{a\gamma_\tau})\frac{d\gamma_\tau}{d\tau}. \end{aligned}$$

Substituting this result into (6.4), after simplification, we get

$$\omega(t,s) = \exp(\frac{-sxe^{bt}}{sae^{bt}\gamma_t + 1})(\frac{1}{sae^{bt}\gamma_t + 1})^{c/a}\Gamma(1 - \frac{c}{a}; \frac{x}{a\gamma_t(sae^{bt}\gamma_t + 1)}).$$

The next step is to perform an inverse Laplace transform with respect to s. To this end, let

$$\begin{aligned} A &= \frac{x}{a\gamma_t}, \\ z &= sae^{bt}\gamma_t + 1. \end{aligned}$$

One can verify that equation (6.5) in [16] is still valid after these substitutions. The quantity $\omega(t,s)$ can now be rewritten as

$$\frac{1}{\Gamma(1-\frac{c}{a})}e^{-A}A^{1-\frac{c}{a}}\int_0^1 (1-\tau)^{-c/a}e^{\frac{A\tau}{z}}z^{-1}\, d\tau.$$

Using the fact that Laplace Transform of $I_0(z(A\tau Y)^{1/2})$ is $e^{\frac{A\tau}{z}}z^{-1}$, we have

$$u(t,Y) = (\frac{1}{ae^{bt}\gamma_t})(\frac{xe^{bt}}{Y})^{\frac{-c+a}{2a}} \exp\{-\frac{(Y + xe^{bt})}{ae^{bt}\gamma_t}\}I_{1-\frac{c}{a}}[\frac{2}{a\gamma_t}(e^{-bt}xY)^{1/2}],$$

where $I_\lambda(\cdot)$ is the first type Bessel function with order λ, which is defined as

$$I_\lambda(\cdot) = \sum_{k=0}^\infty \frac{(\cdot/2)^{2k+\lambda}}{k!\Gamma(k+1+\lambda)}.$$

This density function is then used to find the solution of P at (t_0, X_0) by means of the identity,

$$P(t_0,X_0) = e^{rt_0}Q(t_0,X_0) = e^{-r(T-t_0)}E[\max(Y_T^{\frac{1}{2-\beta}} - K, 0)].$$

After direct calculations, we have

$$\begin{aligned} P(t_0,X_0) &= e^{-r(T-t_0)}\int_{K^{2-\beta}}^\infty (y^{\frac{1}{2-\beta}} - K)u(T,y)\, dy \\ &= e^{-r(T-t_0)}\sum_{r=0}^\infty e^{(r-\delta)T}x^{\frac{1}{2-\beta}}\frac{1}{r!}e^{-\frac{x}{a\gamma_T}}(\frac{x}{a\gamma_T})^r G(r+1+\frac{1}{2-\beta}, \frac{K^{2-\beta}}{ae^{bT}\gamma_T}) \\ &\quad - Ke^{-r(T-t_0)}\sum_{r=0}^\infty \frac{1}{\Gamma(r+1-\frac{1}{2-\beta})}e^{-\frac{x}{a\gamma_T}}(\frac{x}{a\gamma_T})^{r+\frac{1}{2-\beta}}G(r+1, \frac{K^{2-\beta}}{ae^{bT}\gamma_T}). \end{aligned}$$

Putting $x = e^{-bt_0} X_0^{2-\beta}$,

$$x^{\frac{1}{2-\beta}} = e^{-\frac{bt_0}{2-\beta}} X_0 = e^{-(r-\delta)t_0} X_0.$$

So the option pricing formula is

$$P(t_0, X_0) = e^{-\delta(T-t_0)} X_0 \sum_{r=0}^{\infty} \frac{1}{r!} e^{-\frac{x}{a\gamma_T}} (\frac{x}{a\gamma_T})^r G(r+1+\frac{1}{2-\beta}, \frac{K^{2-\beta}}{ae^{bT}\gamma_T})$$
$$- Ke^{-r(T-t_0)} \sum_{r=0}^{\infty} \frac{1}{\Gamma(r+1-\frac{1}{2-\beta})} e^{-\frac{x}{a\gamma_T}} (\frac{x}{a\gamma_T})^{r+\frac{1}{2-\beta}} G(r+1, \frac{K^{2-\beta}}{ae^{bT}\gamma_T}).$$

□

This formula is similar to the classical one, which is obtained by replacing the term γ_T by the term $\frac{1}{2b}(e^{-bt_0} - e^{-bT})$. As these two terms do not depend on the strike price, the implied volatility pattern is the same as the classical CEV model. Consequently, the fractional CEV model also accounts for the volatility skewness observed in practice.

References

[1] Bender, C. (2003). An S-transform approach to integration with respect to a fractional Brownian motion. *Bernoulli* **9** 955–983.

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

[3] Black, F. (1975). Fact and fantasy of using options. *Financial Analysts Journal* **31** 36–72.

[4] Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the Meetings of the American Statistical Association, Business and Economics Statistics Division*, pp. 177–181.

[5] Cox, J. (1996). The constant elasticity of variance option pricing model. *Journal of Portfolio Management* **23** 15–17.

[6] Cox, J. and Ross, S. (1976). The valuation of options for alternative stochastic processes. *J. of Financial Economics* **3** 145–166.

[7] Cutland, N. J., Kopp, P. E. and Willinger, W. (1995). Stock price returns and the Joseph effect: A fractional version of the Black-Scholes model. *Seminar on Stochastic Analysis, Random Fields and Applications, Progr. Probab.* **36** 327–351.

[8] Dai, W. and Heydes, C. C. (1996). Itô formula with respect to fractional Brownian motion and its application. *Journal of Applied Math. and Stoc. Analysis* **9** 439–448.

[9] Decreusefond, L. and Üstünel, A. S. (1999). Stochastic analysis for fractional Brownian motion. *Potential Analysis* **10** 177–214.

[10] Delbaen, F. (1992). Representing martingale measures when asset prices are continuous and bounded. *Mathematical Finance* **2** 107–130

[11] Delbaen, F. and Schacherrnayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Math. Ann.* **300** 463–520.

[12] Duncan, T.E., Hu, Y. and Pasik-Duncan, B. (2000). Stochastic calculus for fractional Brownian motion I: Theory. *SIAM J. Control and Optimization* **38** 582–612.

[13] Embrechts, P. and Maejima, M. (2002). *Self-similar Process*. Princeton University Press, Chichester.

[14] Elliott, R. J. and van de Hoek, J. (2003). A general fractional white noise theory and applications to finance. *Math. Finance* **13** 301–330.

[15] Feller, W. (1951). Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 227–246.

[16] Feller, W. (1951). Two singular diffusion problems. *Annals of Mathematics* **54** 173–182.

[17] Hida, T. (1975). *Analysis of Brownian Functionals. Carleton Math. Lect. Notes* **13**. Carleton University, Ottawa.

[18] Holden, H., Øksendal, B., Ubøe, J. and Zhang, T. (1996). *Stochastic Partial Differential Equations*. Birkhäuser, Boston.

[19] Hu, Y. and Øksendal, B. (2003). Fractional white noise calculus and application to finance. *Infinite Dimensional Analysis, Quantum Probability and Related Topics* **6** 1–32.

[20] Huang, Z. and Yan, J. (2000). *Introduction to Infinite Dimensional Stochastic Analysis*. Science Press/ Kluwer Academic Publishers.

[21] Kolmogorov, A. N. (1931). Analytical methods in probability theory. *Selected works of A.N. Kolmogorov*, Vol 2, edited by A.N. Shiryaev, pp. 62–107.

[22] Kuo, H.-H. (1996). *White Noise Distribution Theory*. CRC Press, Boca Raton, FL.

[23] Lo, A. W. and MacKinlay, A. C. (1999). *A Non-Random Walk down Wall Street*. Princeton University Press, New Jersey.

[24] Malliavin, P. and Thalmaiar, A. (2005). *Stochastic Calculus of Variations in Mathematical Finance*. Springer-Verlag, Heidelberg.

[25] Milstein, G. N. (1995). *Numerical integration of stochastic differential equations*. Kluwer Academic Publishers, Boston.

[26] Necula, C. (2002). Option pricing in a fractional Brownian motion environment. Manuscript, Academy of Economic Studies Bucharest.

[27] Nualart, D. (1995). *The Malliavin Calculus and Related Topics*. Springer-Verlag, New York.

[28] Obata, N. (1995). *White Noise Calculus and Fock Space. Lecture Notes in Mathematics* **1577**. Springer-Verlag, Heidelberg.

[29] Rogers, L. C. G. (1997). Arbitrage with fraction Brownian motion. *Mathematical Finance* **7** 95–105.

[30] Samorodnitsky, G. and Taqqu, M. (1994). *Stable non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York.

[31] Zhang, T. S. (1992). Characterizations of white noise test functions and Hida distributions. *Stochastics* **41** 71–87.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 165–172

DOI: 10.1214/074921706000001021

Estimation errors of the Sharpe ratio for long-memory stochastic volatility models

Hwai-Chung Ho[1]

Academia Sinica

Abstract: The Sharpe ratio, which is defined as the ratio of the excess expected return of an investment to its standard deviation, has been widely cited in the financial literature by researchers and practitioners. However, very little attention has been paid to the statistical properties of the estimation of the ratio. Lo (2002) derived the $\sqrt{n}$-normality of the ratio's estimation errors for returns which are iid or stationary with serial correlations, and pointed out that to make inference on the accuracy of the estimation, the serial correlation among the returns needs to be taken into account. In the present paper a class of time series models for returns is introduced to demonstrate that there exists a factor other than the serial correlation of the returns that dominates the asymptotic behavior of the Sharpe ratio statistics. The model under consideration is a linear process whose innovation sequence has summable coefficients and contains a latent volatility component which is long-memory. It is proved that the estimation errors of the ratio are asymptotically normal with a convergence rate slower than $\sqrt{n}$ and that the estimation deviation of the expected return makes no contribution to the limiting distribution.

1. Introduction

An interesting phenomenon observed in many financial time series is that strong evidence of persistent correlation exists in some nonlinear transformation of returns, such as square, logarithm of square, and absolute value, whereas the return series itself behaves almost like white noise. This so-called clustering volatility property has a profound implication. The traditional linear processes such as ARMA models and the mixing conditions of various types that have been widely used to account for the weak-dependence or short-memory properties of stationary processes (see, e.g., [1]) are found inadequate to model the dependence structure of the return process. A great deal of research works have been devoted to looking for proper models that entail the stylized fact mentioned above. The ARCH model proposed by Engle [6] and its various extensions are attempts that have been proved very successful. Recently, models other than ARCH family have been seen to provide better fitting for data with clustering volatility. For instance, Lobato and Savin [11] examine the S&P 500 index series for the period of July 1962 to December 1994 and report that the squared daily returns exhibit the genuine long-memory effect which ARCH process cannot produce (see also [5]). Based on Lobato and Savin's finding, Breidt, Crato and Lima [2] suggest the following long-memory stochastic volatility model (LMSV):

$$r_t = v_t \varepsilon_t, \quad v_t = \delta \exp(x_t), \tag{1.1}$$

[1]Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, e-mail: hcho@stat.sinica.edu.tw

AMS 2000 subject classifications: primary 60G10, 62M10; secondary 60F05.
Keywords and phrases: long memory, stochastic volatility, Sharpe ratio.

where $\delta > 0$, and $\{x_t\}$ is a Gaussian process which exhibits long memory and is independent of the iid sequence $\{\varepsilon_t\}$ with mean zero and variance one. The short-memory version of model (1.1) has been discussed, for example, by Taylor [14], Melino and Turnbull [12] and Harvey et al. [8]. The precise definition of short- or long-memory process is given as follows. A linaer process defined as

$$x_t = \sum_{i=0}^{\infty} a_i z_{t-i}, \tag{1.2}$$

where the z_i are iid random variables (Gaussian or non-Gaussian) having mean 0 and variance one, is called short-memory if the coefficients a_i are summable or long-memory if $a_i \sim Ci^{-\beta}$ with β in $(1/2, 1)$; "$g_n \sim h_n$" signifies $\lim_{n\to\infty} g_n/h_n = 1$. The long-memory process just defined is sometimes also referred to as a fractional differencing (or $I(d)$) process with the memory parameter $d = 1-\beta$ [3]. It can be seen that the LMSV model described in (1.1) and (1.2) exhibits the desirable property that $\{r_t\}$ is white noise and $\{r_t^2\}$ is long-memory. Because of this characteristic property, one needs to be cautious in making statistical inference for the LMSV model if the statistics of interest involve nonlinear transformations. The purpose of this paper is to point out a circumstance under which the estimation statistics based on the LMSV model behave distinctly different from traditional stationary sequences of weak dependence such as the ARMA model with iid innovations.

We use the example of the Sharpe ratio to demonstrate that for the LMSV model the estimation statistics have entirely different asymptotic properties from those of the case where the volatilily is short-memory. Discussions of this and a more general model are presented in Sections 2 and 3, respectively. The paper's main result is formulated in a theorem stated in Section 3 and its proof is given in Section 4.

2. LMSV models: the simple case

The Sharpe raio, which is defined as the ratio of the excess expected return of an investment to its standard deviation, is originally motivated by the mean-variance analysis and the Sharpe-Lintner Captial Asset Pricing Model (Campbell, Lo and MacKinlay [4]) and has become a popular index used to evaluate investment performance and for risk management. Both the expected return and the standard deviation are generally unknown and need to be estimated. Although the ratio is one of the most commonly cited statistics in financial analysis by researchers and practitioners as well, not much attention has been paid to its statistical properties until the work of Lo [10]. Lo [10] points out that to gauge the accuracy of the estimates of the ratio, it is important to take into account the dependence of the returns for it may result significant difference of the limiting variance between iid and non-iid (dependent) returns. For both of the two cases the standard $\sqrt{n}$ central limit theorem is assumed to hold for the ratio's estimates. The LMSV time series is a stationary martingale difference sequence bearing strong dependence in the latent component of volatility. The partial sums of the sequence itself and of the sequence after a certain transformation is applied may have entirely different asymptotic behaviors. Below we show that for the LMSV model, the Sharpe ratio statistic is asymptotically normal but converges to the true ratio at a rate slower than $\sqrt{n}$. Furthermore, while the ratio's statistics involve the estimates of the expected return and the standard deviation, it turns out that only the estimation errors of the latter contribute to the limit distribution as opposed to the case of short-memory volatility where neither of the two estimates is asymptotically negligible.

Let the returns $\{r_t\}$ be model as in (1.1) and (1.2) with long-memory $x_t = \sum_{i=0}^{\infty} a_i z_{t-i}$, where z_i are iid random variables having mean 0 and variance 1 and the coefficients a_i are such that $a_i \sim C \cdot i^{-\beta}$ with β in $(1/2, 1)$. Denote by $\sigma^2 = Er_t^2$. For the observed returns $\{r_1, \ldots, r_n\}$, we define

$$\hat{\mu} = n^{-1} \sum_{t=1}^{n} r_t, \quad \hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} (r_t - \hat{\mu})^2,$$

and the Sharpe ratio statistics

$$\hat{SR} = \frac{\hat{\mu} - r_f}{\hat{\sigma}},$$

where r_f is a fixed risk-free interest rate assumed to be positive. Using the δ-method, we have

$$\hat{SR} - SR = \frac{\hat{\mu}}{\sigma} + \frac{r_f(\hat{\sigma}^2 - \sigma^2)}{2\sigma^3} + O_p((\hat{\sigma}^2 - \sigma^2)^2).$$

Also write

$$\hat{\sigma}^2 - \sigma^2 = n^{-1} \sum_{t=1}^{n} v_t^2(\varepsilon_t^2 - 1) + n^{-1} \sum_{t=1}^{n} (v_t^2 - \sigma^2) - \hat{\mu}^2.$$

To derive the asymptotic distribution, we first compute the variance of $\hat{\sigma}^2 - \sigma^2$. Note that

$$\text{(2.1)} \qquad \operatorname{var}(n^{-1} \sum_{t=1}^{n} v_t^2(\varepsilon_t^2 - 1)) = O(n^{-1}) \quad \text{and} \quad \operatorname{var}(\hat{\mu}) = O(n^{-1}),$$

since both $\{v_t^2(\varepsilon_t^2 - 1)\}$ and $\{v_t \varepsilon_t\}$ are sequences of martingale differences. For $\sum_{t=1}^{n} (v_t^2 - \sigma^2)$, we use the results obtained by Ho and Hsing [9]. Let $F(\cdot)$ be the common distribution function of the x_t. Denote by

$$K_\infty(y) = e^{2y} \int e^{2x} dF(x).$$

Then by Theorem 3.1 and Corollary 3.3 of Ho and Hsing [9],

$$\text{(2.2)} \qquad n^{\beta-3/2}\{\sum_{t=1}^{n} (v_t^2 - \sigma^2)\} = \delta^2 K_\infty^{(1)}(0) n^{\beta-3/2}\{\sum_{t=1}^{n} x_t\} + o_p(1)$$
$$\xrightarrow{d} 2\sigma^2 \cdot N(0, \xi^2)$$

with

$$\xi^2 = C^2 \frac{\int_0^\infty (x^2 + x)^{-\beta} dx}{2(1-\beta)(3/2+\beta)} \cdot \int_{-\infty}^{1} \{\int_0^1 [(v-u)^+]^{-\beta} dv\} du.$$

Combining (2.1) and (2.2) gives

$$\text{(2.3)} \qquad n^{\beta-1/2}(\hat{SR} - SR) = \frac{r_f}{2\sigma^3} n^{\beta-1/2}(\hat{\sigma}^2 - \sigma^2) + o_p(1)$$
$$\xrightarrow{d} r_f \sigma^{-1} \cdot N(0, \xi^2),$$

If x_t is short-memory in the sense as specified before that

$$x_t = \sum_{i=0}^{\infty} a_i z_{t-i} \quad \text{with} \quad \sum_{i=1}^{\infty} |a_i| < \infty,$$

then the usual $\sqrt{n}$ central limit theorem will hold for $\sqrt{n}(\hat{SR} - SR)$. The proof of this will be covered in the next subsection as a special case of a more general model.

3. Linear processes of LMSV models

We now focus on the linear process with its innovations being a LMSV sequence. Specifically, define

$$y_t = \sum_{j=0}^{\infty} b_j r_{t-j} \quad \text{with} \quad \sum_{j} |b_j| < \infty, \tag{3.1}$$

where r_t is modeled in (1.1) and (1.2) with $\delta = 1$. Denote by σ_y^2 the variance of the y_t. The Sharpe ratio now is $SR = r_f/\sigma_y$ and its corresponding estimator is

$$\hat{SR}_y = \frac{W_n - r_f}{\hat{\sigma}_y}, \tag{3.2}$$

where

$$W_n = n^{-1} \sum_{t=1}^{n} y_t, \quad \hat{\sigma}_y = (n^{-1} \sum_{t=1}^{n} (y_t - W_n)^2)^{1/2}.$$

From now on we assume that there is a positive constant K such that for any $\eta > 0$

$$Ee^{\eta x_1} \le e^{K\eta^2}. \tag{3.3}$$

As can be seen later in the proof we only need a sufficiently large constant K. Using a stronger condition here is merely for the ease of presentation.

Theorem. *For the model defined in (3.1), assume condition (3.3) holds.*

(i) *Suppose x_t is short-memory, that is, $\sum_{i=0}^{\infty} |a_i| < \infty$. Assume $E\varepsilon_1^3 = 0$, then*

$$\sqrt{n}(\hat{SR} - SR) \xrightarrow{d} N(0, \xi_1^2) \tag{3.4}$$

for some constant ξ_1.

(ii) *If x_t is long-memory with the coefficients satisfying that $a_i \sim Ci^{-\beta}$ for $\beta \in (1/2, 1)$, then*

$$n^{\beta - 3/2}(\hat{SR} - SR) \xrightarrow{d} 2 \int e^{2x} dF(x) N(0, \xi_2^2) \tag{3.5}$$

for some constant ξ_2.

The limiting variances, ξ_1^2 and ξ_2^2, given in (3.4) and (3.5) above will be derived in the proof of the theorem. Both ξ_1^2 and ξ_2^2 depend on the linear filter $\{b_j\}$ and some parameters of the laten process $\{x_t\}$. It is a very challanging problem to estimate the two quantities. For part (ii) of the Theorem, if the distribution function $F(\cdot)$

of x_t is known, then one can use the sampling window method proposed in [7] and [13] to consistently estimate ξ_1^2 and ξ_2^2. As for the short-memory case of part (i) of the Theorem, no existing results in the literature cover this case unless a certain kind of weak dependence is assumed. With only the summability condition on $\{a_j\}$ one needs to develop some new theory to support the use of the resampling scheme mentioned above.

Proof of Theorem. (i) Define

$$x_{t,m} = \sum_{i=0}^{m-1} a_i \varepsilon_{t-i}, \quad \tilde{x}_{t,m} = \sum_{i=m}^{\infty} a_i \varepsilon_{t-i}, \quad r_{t,m} = e^{x_{t,m}} \varepsilon_t, \quad y_{t,m} = \sum_{j=0}^{m-1} b_j r_{t-j,m},$$

$$W_{n,m} = n^{-1} \sum_{t=1}^{n} y_{t,m}.$$

Since $y_{t,m}$'s are $2m$-dependent, as $n \to \infty$,

$$\sqrt{n} W_{n,m} \xrightarrow{d} N(0, \lambda_m^2), \tag{3.6}$$

where

$$\begin{aligned} \lambda_m^2 &= \lim_{n\to\infty} n^{-1} \mathrm{var}\Big(\sum_{t=1}^{n} y_{t,m}\Big) \\ &= \delta^2 E e^{2x_{1,m}} \Big(\sum_{j=0}^{m} b_j^2 + 2 \sum_{k=1}^{\infty} \sum_{j=0}^{m} b_j b_{j+k}\Big). \end{aligned}$$

Write

$$\begin{aligned} \sqrt{n}(W_n - W_{n,m}) &= n^{-1/2} \sum_{t=1}^{n} (y_t - y_{t,m}) \\ &= n^{-1/2} \sum_{t=1}^{n} \sum_{j=0}^{m-1} b_j (r_{t-j} - r_{t-j,m}) + n^{-1/2} \sum_{t=1}^{m} \sum_{j=m}^{\infty} b_j r_{t-j} \\ &\equiv C_{n,m} + D_{n,m}. \end{aligned}$$

Then

$$E C_{n,m}^2 = \delta^2 E e^{2x_{1,m}} (e^{\tilde{x}_{1,m}} - 1) \sum_{k=-n-1}^{n-1} (1 - \frac{|k|}{n}) (\sum_{j=0}^{m-1} b_j b_{j+k}).$$

By using the elementary inequality $|e^x - 1| \le e|x|$, $|x| \le 1$, and the Chebyshev inequality, we have

$$\begin{aligned} E(e^{\tilde{x}_{1,m}} - 1)^2 &= E(e^{\tilde{x}_{1,m}} - 1)^2 I\{\tilde{x}_{1,m} \le 1\} + E(e^{\tilde{x}_{1,m}} - 1)^2 I\{\tilde{x}_{1,m} > 1\} \\ &\le e(E\tilde{x}_{1,m}^2) + (E(e^{\tilde{x}_{1,m}} - 1)^4)^{1/2} (E\tilde{x}_{1,m}^2)^{1/2}. \end{aligned}$$

Because, by assumption (3.3), $E e^{4x_{1,m}}$ is bounded in m, we have

$$E(e^{\tilde{x}_{1,m}} - 1)^2 \to 0 \quad \text{as} \quad m \to \infty. \tag{3.7}$$

This and $\sum_j^{\infty} |b_j| < \infty$ jointly imply

$$\lim_{m\to\infty} \lim_{n\to\infty} E C_{n,m}^2 = 0. \tag{3.8}$$

Similarly,

$$\lim_{m\to\infty}\lim_{n\to\infty} ED_{n,m}^2 = 0. \tag{3.9}$$

From (3.6), (3.8) and (3.9) it follows that

$$\sqrt{n}W_n \to N(0,\lambda^2), \tag{3.10}$$

where

$$\lambda^2 = \lim_{m\to\infty}\lambda_m^2 = \sigma^2(\sum_{j=0}^{\infty} b_j^2 + \sum_{k=1}^{\infty}\sum_{j=0}^{\infty} b_j b_{j+k}).$$

We now derive the limiting distribution for $\sqrt{n}(\hat{\sigma}_y^2 - \sigma_y^2)$. Write

$$\begin{aligned}\sqrt{n}(\hat{\sigma}_y^2 - \sigma_y^2) &= \delta^2 n^{-1/2}\sum_{t=1}^{n}\sum_{j=0}^{\infty} b_j^2 e^{2x_{t-j}}(\varepsilon_{t-j}^2 - 1) + \delta^2 n^{-1/2}\sum_{t=1}^{n}\sum_{j=0}^{\infty}(e^{2x_{t-j}} - \sigma_y^2)\\ &\quad + n^{-1/2}\sum_{t=1}^{n}\sum_{i\neq j} b_i b_j r_{t-i} r_{t-j}\\ &\equiv V_{n,1} + V_{n,2} + V_{n,3}. \end{aligned}\tag{3.11}$$

By the same m- truncation argument as used in proving (3.8) one can show that $V_{n,1}$, $V_{n,2}$ and $V_{n,3}$ are asymptotically normal and independent, that is, as $n\to\infty$,

$$V_{n,1} + +V_{n,2} + V_{n,3} \xrightarrow{d} N(0, g^2), \tag{3.12}$$

where g^2 is the sum of the limiting variances of $V_{n,1}$, $V_{n,2}$ and $V_{n,3}$. Because x_t may be non-Gaussian, the analytic form of the covariance function of $\{e^{2x_t} - \sigma_y^2\}$ and consequently of the limiting variance of $V_{n,2}$, which equals to

$$\delta^2 \lim_{m\to\infty}\limsup_{n\to\infty} n^{-1}\mathrm{var}(\sum_{t=1}^{n}(e^{2x_{t,m}} - \sigma_y^2)),$$

is not available. However, the exact formulas of limiting variances for $V_{n,1}$and $V_{n,3}$ can be found as follows.

$$\lim_{n\to\infty}\mathrm{var}(V_{n,1}) = \delta^4[Ee^{4x_1}][E(\varepsilon_1^2 - 1)^2][\sum_{j=0}^{\infty} b_j^4 + 2\sum_{k=1}^{\infty}\sum_{j=0}^{\infty} b_j^2 b_{j+k}^2],$$

$$\lim_{n\to\infty}\mathrm{var}(V_{n,3}) = \delta^4[Ee^{2x_1}]^2[\sum_{i\neq j} b_i^2 b_j^2 + \sum_{k=1}^{\infty}\sum_{i\neq j} b_i b_{i+k} b_j b_{j+k}].$$

Note that the assumption $E\varepsilon_1^3 = 0$ is used to prove that $V_{n,2}$ is asymptotically independent with $V_{n,1}$ and $V_{n,3}$. The limit results of (3.10)and (3.12) imply

$$\sqrt{n}(\hat{SR} - SR) \xrightarrow{d} N(0, \xi_1^2)$$

with $\xi_1^2 = \lambda^2 + r_f^2(4\sigma_y^6)^{-1}g^2$. Hence (2.4) holds.
(ii) Because $\{r_t\}$ is a sequence of martingale differences, we have

$$\mathrm{var}(W_n) = O(1/n). \tag{3.13}$$

Similarly, for $V_{n,1}$ and $V_{n,3}$ defined in (3.11),

$$\text{var}(V_{n,1}) = O(1), \quad \text{var}(V_{n,3}) = O(1). \tag{3.14}$$

To compute the variance of $V_{n,2}$, define $y'_t = \sum_{j=0}^{\infty} b_j x_{t-j}$. Then y'_t can be rewritten as

$$y'_t = \sum_{j=0}^{\infty} z_{t-j} B_j,$$

where

$$B_j = \sum_{i=0}^{j} b_i a_{j-i}.$$

As $j \to \infty$, since $a_j \sim Cj^{-\beta}$,

$$B_j \sim C_1 j^{-\beta},$$

where $C_1 = C(\sum_{i=0}^{\infty} b_i)$. Then, as $k \to \infty$,

$$\sum_{j=0}^{\infty} B_j B_{j+k} \sim C_1^2 \int x^{-\beta}(1+x)^{-\beta} dx \cdot k^{-2\beta+1},$$

implying that

$$Ey'_t y'_{t+k} = \sum_{j=0}^{\infty} B_j B_{j+k} \sim C_1^2 \int x^{-\beta}(1+x)^{-\beta} dx \cdot k^{-2\beta+1}.$$

In other words, $\{y'_t\}$ is also a linear long-memory process having the same memory parameter as that of x_t. Therefore, similar to (2.2),

$$n^{\beta-3/2} \sum_{t=1}^{n} y'_t \xrightarrow{d} N(0, \xi_2^2) \tag{3.15}$$

with

$$\xi_2^2 = \frac{C_1^2 \int_0^{\infty} (x^2+x)^{-\beta} dx}{2(1-\beta)(3/2+\beta)} \cdot \int_{-\infty}^{1} \{ \int_0^1 [(v-u)^+]^{-\beta} dv \} du.$$

As noted before in (2.2) that

$$n^{\beta-3/2} \{ \sum_{t=1}^{n} (e^{2x_t} - \sigma^2) \} = 2 \int e^{2x} dF(x) \cdot (n^{\beta-3/2} \{ \sum_{t=1}^{n} x_t \}) + o_p(1).$$

From this and (3.15), we have, as $n \to \infty$,

$$\begin{aligned} n^{\beta-3/2} \{ \sum_{t=1}^{n} \sum_{j=0}^{\infty} b_j (e^{2x_{t-j}} - \sigma_y^2) \} &= 2 \int e^{2x} dF(x) (n^{\beta-3/2} \sum_{t=1}^{n} y'_t) + o_p(1) \\ &\xrightarrow{d} 2 \int e^{2x} dF(x) \cdot N(0, \xi_2^2). \end{aligned} \tag{3.16}$$

Summarizing (3.13), (3.14) and (3.16) gives

$$n^{\beta-3/2}(\hat{SR} - SR) \xrightarrow{d} 2 \int e^{2x} dF(x) N(0, \xi_2^2).$$

The proof is completed. □

References

[1] Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*. E. Eberlein and M. S. Taqqu (eds.). Birkhäuser, Boston, pp. 162–192.

[2] Breidt, F. J., Crato, N., and Lima, P. (1998). The detection and estimation of long memory in the stochastic volatility. *Journal of Econometrics* **83** 325–348.

[3] Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer, New York.

[4] Campbell, J., Lo, A. and MacKinlay, C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

[5] Ding, Z., Granger, C. W. J., and Engle, R. F. (1993). A long memory property of stock returns and a new model. *Journal of Empirical Finance* **1** 83–106.

[6] Engle, R. (1982). Autoregressive conditional hetroskedasticity with estimates of the variance of U.K. inflation. *Econometrika* **50** 987–1008.

[7] Hall, P., Jing, B. Y. and Lahiri, S. N. (1998). On the sampling window method for long-range dependent data. *Statistica Sinica* **8** 1189–1204.

[8] Harvey, A. C., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *Review of Financial Studies* **61** 247–264.

[9] Ho, H.-C. and Hsing, T. (1997). Limit theorems for functional of moving averages. *Ann. Probab.* **25** 1636–1669.

[10] Lo, A. W. (2002). The statistics of Sharpe ratio. *Financial Analysts Journal* **58** 36–52.

[11] Lobato, I. N. and Savin, N. E. (1998). Real and spurious long-memory properties of stock-market data. *Journal of Business and Economics Statistics* **16** 261–277.

[12] Melino, A. and Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* **45** 239–265.

[13] Nordman, D. J. and Lahiri, S. N. (2005). Validity of the sampling window mwthod for long-range dependent linear processes. *Econometric Theory* **21** 1087–1111.

[14] Taylor, S. J. (1986). *Modeling Financial Time Series*. John Wiley & Sons, New York.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 173–192

DOI: 10.1214/074921706000001030

Cowles commission structural equation approach in light of nonstationary time series analysis

Cheng Hsiao[1]

University of Southern California

Abstract: We review the advancement of nonstationary time series analysis from the perspective of Cowles Commission structural equation approach. We argue that despite the rich repertoire nonstationary time series analysis provides to analyze how do variables respond dynamically to shocks through the decomposition of a dynamic system into long-run and short-run relations, nonstationarity does not invalid the classical concerns of structural equation modeling — identification and simultaneity bias. The same rank condition for identification holds for stationary and nonstationary data and some sort of instrumental variable estimators will have to be employed to yield consistency. However, nonstationarity does raise issues of inference if the rank of cointegration or direction of nonstationarity is not known *a priori*. The usual test statistics may not be chi-square distributed because of the presence of unit roots distributions. Classical instrumental variable estimators have to be modified to ensure valid inference.

1. Introduction

Let $\{\underset{\sim}{w}_t\}$ be a sequence of time series observations of random variables. Multivariate vector autoregressive model (VAR) has been suggested as a useful tool to summarize the information contained in the data and to generate predictions (e.g. Hsiao [21, 22], Sims [50]). These models treat all variables as joint dependent and treat $\underset{\sim}{w}_t$ as a function of its past values, $\underset{\sim}{w}_{t-j}$. On the other hand, Cowles Commission approach assumes each equation in the system describes a behavioral or technological relations. An essential element of the Cowles Commission approach is to decompose $\underset{\sim}{w}_t$ into G endogenous variables, $\underset{\sim}{y}_t$, and K exogenous variables, $\underset{\sim}{x}_t, \underset{\sim}{w}_t' = (\underset{\sim}{y}_t', \underset{\sim}{x}_t'), G + K = m$. The value of endogenous variables $\underset{\sim}{y}_t$ are determined by the simultaneous interaction of the behavioral, technological or institutional relations in the model given the value of the exogenous variables, $\underset{\sim}{x}_t$, and shock of the system (say, $\underset{\sim}{\epsilon}_t$). The value of $\underset{\sim}{x}_t$ is assumed to be determined by the forces outside of the model (e.g. Koopmans and Hood [19]). The Cowles Commission structural equation approach is also referred as a structural equations model (SEM). It has wide applications in education, psychology and econometrics, etc. (e.g. Browne and Arminger [6], Hood and Koopmans [19], Muthen [39, 40], Yuan and Bentler [59]). In this paper we will only focus on the aspects related to the time series analysis of a SEM.

Since the observed data can only provide information on conditional distribution of $\underset{\sim}{y}_t$ given past values of $\underset{\sim}{y}_{t-j}$ and current and past values of $\underset{\sim}{x}_{t-j}$, there is an issue of if it is possible to infer from the data the true data generating process for the SEMs, which is referred to as an *identification* issue. Another issue for the SEMs is because

[1]Department of Economics, University of Southern California, 3620 S. Vermont Ave. KAP300, Los Angeles, CA 90089, e-mail: chsiao@usc.edu

of the joint dependency of y_t, the regressors of an equation are correlated with the error (shock) of an equation which violates the condition for the regression method to be consistent. This is referred to as *simultaneity bias* issue. The theory and statistical properties of SEMs are well developed for stationary data (e.g. Amemiya [2], Intriligator, Boskin and Hsiao [30]).

Nelson and Plosser [41] have shown that many economic and financial data contain unit roots, namely, most are integrated of order 1 or 2, I(1) or I(2). Theories for the time series analysis with unit roots have been derived by Anderson [4], Chan and Wei [7], Johansen [31, 32], Phillips [45], Phillips and Durlauf [46], Sims, Stock and Watson [51], Tiao and Tsay [57], etc. Among the major findings are that (i) w_t may be cointegrated in the sense that a linear combination of $I(d)$ variables may be of order $I(d-c)$, where d and c are positive numbers, say 1 (Granger and Weiss [14], Engle and Granger [11], Tiao and Box [54]); (ii) "Since these models (VAR) don't dichotomize variables into "endogenous" and "exogenous," the exclusion restrictions used to identify traditional simultaneous equations models make little sense" (Watson [58]); (iii) Time series regressions with integrated variables can behave very differently from those with stationary variables. Some of the estimated coefficients converge to their true values at the speed of $\sqrt{T}$ and are asymptotically normally distributed. Some converge to the true values at the speed of T but have non-normal asymptotic distribution, and are asymptotically biased. Hence the Wald test statistics under the null may not be approximated by chi-square distributions (Chan and Wei [7], Sims, Stock and Watson [51], Tsay and Tiao [57]); (iv) Even though the $I(1)$ regressors may be correlated with the errors, the least squares regression consistently estimates the cointegrating relation, hence the simultaneity bias issues may be ignored (Phillips and Durlauf [46], Stock [52]).

In this paper we hope to review the recent advances in nonstationary time series analysis from the perspective of Cowles Commission Structural equation approach. In section 2 we discuss the relationships between a vector autoregressive model (VAR), a structural vector autoregressive model (SVAR), and Cowles Commission structural equations model (SEM). Section 3 discusses issues of estimating VAR with integrated variables. Section 4 discusses the least squares and instrumental variable estimators, in particular, the two stage least squares estimator (2SLS) for a SVAR. Section 5 discusses the modified and lag order augmented 2SLS estimators for SVAR. Conclusions are in Section 6.

2. Vector autoregression, structural vector autoregression and structural equations model

For ease of exposition, we shall assume that all elements of w_t are I(1) processes. We assume that w_t are generated by the following p-th order structural vector autoregressive process without intercept terms:[1]

$$A(L)w_t = \epsilon_t \tag{2.1}$$

where $A(L) = A_0 + A_1 L + A_2 L^2 + \cdots + A_p L^p$. We assume that initial observations $w_0, w_{-1}, \ldots, w_{-p}$ are available and

A.1: A_0 is nonsingular and $A_0 \neq I_m$, where I_m denotes an m rowed identity matrix.
A.2: The roots of $|A(L)| = 0$ are either 1 or outside the unit circle.

[1] The introduciton of intercept terms complicates algebraic manipulation without changing the basic message. For detail, see [28].

A.3: The $m\times 1$ error or innovation vector $\underset{\sim}{\epsilon}_t$ is independently, identically distributed (i.i.d.) with mean zero, nonsingular covariance matrix $\Sigma_{\epsilon\epsilon}$ and finite fourth cumulants.

Premultiplying A_0^{-1} to (2.1) yields the conventional VAR model of Johansen [31, 32], Phillips [45], Sims [50], Sims, Stock and Watson [51], Tsay and Tiao [57], etc.,

$$\underset{\sim}{w}_t = \Pi_1 \underset{\sim}{w}_{t-1} + \cdots + \Pi_p \underset{\sim}{w}_{t-p} + \underset{\sim}{v}_t, \tag{2.2}$$

where $\Pi_j = -A_0^{-1}A_j, j = 1, \ldots, p$, and $\underset{\sim}{v}_t = A_0^{-1}\underset{\sim}{\epsilon}_t$. The difference between (2.1) and (2.2) is that each equation in the former is supposed to describe a behavioral or technological relation while the latter is a *reduced form* relation. Eq. (2.2) is useful for generating prediction, but cannot be used for structural or policy analysis. For instance, $w_{1t}, w_{2t}, w_{3t}, w_{4t}$ may denote the price and quantity of a product, per capita income and raw material price, respectively. The first and second equations describe a demand relation which has quantity inversely related to price and positively related to income, and a supply relation which has price positively related to quantity and raw material price, respectively. Only (2.1) can provide information on demand and supply price elasticities but not (2.2). Equation (2.2) can only yield expected value of price and quantity given past $\underset{\sim}{w}_{t-j}$.

Let $A = [A_0, A_1, \ldots, A_p]$ and define a $(p+1)m$-dimensional nonsingular matrix M as

$$M = \begin{bmatrix} I_m & I_m & \ldots & I_m \\ \underset{\sim}{0} & I_m & \ldots & I_m \\ \underset{\sim}{0} & \underset{\sim}{0} & \ldots & I_m \\ . & . & \ldots & \\ \underset{\sim}{0} & . & .\underset{\sim}{0}. & I_m \end{bmatrix}. \tag{2.3}$$

Postmultiplying A by M yields an error-correction representation of (2.1),

$$\sum_{j=0}^{p-1} A_j^* \bigtriangledown \underset{\sim}{w}_{t-j} + A_p^* \underset{\sim}{w}_{t-p} = \underset{\sim}{\epsilon}_t, \tag{2.4}$$

where $\bigtriangledown = (1-L), A_j^* = \sum_{\ell=0}^{j} A_\ell, j = 0, 1, \ldots, p$. Let $A^* = [A_1^*, \ldots, A_p^*] = [\tilde{A}_1^*, A_p^*]$, then $A^* = AM$. The coefficient matrices $\tilde{A}_1^*$ and A_p^* provide the implied *short-run* dynamics and *long-run* relations of the system (2.1) as defined in [26].[2]

Similarly, we can post-multiply (2.2) by M to yield an *error-correction* representation of the reduced form (2.2)

$$\bigtriangledown \underset{\sim}{w}_t = \Pi_1^* \bigtriangledown \underset{\sim}{w}_{t-1} + \cdots + \Pi_{p-1}^* \bigtriangledown \underset{\sim}{w}_{t-p+1} + \Pi_p^* \underset{\sim}{w}_{t-p} + \underset{\sim}{v}_t, \tag{2.5}$$

where $\Pi_j = \sum_{i=1}^{j} \Pi_i - I_m$.

In this paper we are concerned with statistical inference of (2.1). If the roots of $|A(L)| = 0$ are all outside the unit circle, $\underset{\sim}{w}_t$ is stationary. It is well known that the least squares estimator (LS) is inconsistent. The 2SLS and 3SLS using lagged $\underset{\sim}{w}_t$ as instruments are consistent and asymptotically normally distributed (e.g. Amemiya [2], Malinvaud [38]). Therefore, we shall assume that at least one root of $|A(L)| = 0$

[2]The long-run and short-run dichotomization defined here is derived from (2.1). They are different from the those implied by Granger and Lin [13], Johansen [31, 32] or Pesaran, Shin and Smith [43], etc.

is equal to 1. More specifically,[3]

A4:(a) $A_p^* = \underset{\sim}{\alpha}\underset{\sim}{\beta}'$ (or $\Pi_p^* = \underset{\sim}{\alpha}^*\underset{\sim}{\beta}^{*'}$) where $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ (or $\underset{\sim}{\alpha}^*$ and $\underset{\sim}{\beta}^*$) are $m \times r$ matrices of full column rank $r, 0 \leq r \leq m-1$

(b) $\underset{\sim}{\alpha}_\perp' J \underset{\sim}{\beta}_\perp$ or $(\underset{\sim}{\alpha}_\perp^{*'} J^* \underset{\sim}{\beta}_\perp^*)$ is nonsingular, where $J = \sum_{j=0}^{p-1} A_j^*$, (or $J^* = \sum_{j=0}^{p-1} \Pi_j^*$), $\underset{\sim}{\alpha}_\perp$ and $\underset{\sim}{\beta}_\perp$ (or $\underset{\sim}{\alpha}_\perp^*$ and $\underset{\sim}{\beta}_\perp^*$) are $m \times (m-r)$ matrices of full column rank such that $\underset{\sim}{\alpha}_\perp'\underset{\sim}{\alpha} = \underset{\sim}{0} = \underset{\sim}{\beta}_\perp'\underset{\sim}{\beta}$, (or $\underset{\sim}{\alpha}_\perp^{*'}\underset{\sim}{\alpha}^* = \underset{\sim}{0} = \underset{\sim}{\beta}_\perp^{*'}\underset{\sim}{\beta}$) (If $r = 0$, then we take $\underset{\sim}{\alpha}_\perp = I_m = \underset{\sim}{\beta}_\perp$.)

Under A1-A4, $\underset{\sim}{w}_t$ has r *cointegrating* vectors (the columns of β) and $m-r$ unit roots. As shown by Johansen [31, 32] and Toda and Phillips [56] that A4 ensures that the *Granger representation theorem* (Engle and Granger [11]) applies, so that $\bigtriangledown \underset{\sim}{w}_t$ is stationary, $\underset{\sim}{\beta}'\underset{\sim}{w}_t$ is stationary, and $\underset{\sim}{w}_t$ is an I(1) process when $r < m$.

The cointegrating vectors $\underset{\sim}{\beta}$ provide information on the "long-run" or "equilibrium" state in which a dynamic system tends to converge over time after any of the variables in the system being perturbed by a shock, $\underset{\sim}{\alpha}$ transmits the deviation from such long-run relations, $\underset{\sim}{e}_t = \underset{\sim}{\beta}'\underset{\sim}{w}_t$, into each of $\underset{\sim}{w}_t$, and $\tilde{A}_1^*$ provides information on how soon such "equilibrium" is restored. In economics, the existence of long-run relationships and strength of attraction to such a state depends on the actions of a market or on government intervention. In this sense, the concept of *cointegration* has been applied in a variety of economic models including the relationships between capital and output; real wages and labor productivity; nominal exchange rate and relative prices, consumption and disposable income, long- and short-term interest rates, money velocity and interest rates, price of shares and dividends, production and sales, etc. (e.g. Banerjee, Dolado, Galbraith and Hendry [5], Hsiao, Shen and Fujiki [29], King, Plosser, Stock and Watson [33]).

Since the data only provide information of the conditional density of $\underset{\sim}{w}_t$ given past values of $\underset{\sim}{w}_{t-j}, j = 1, \ldots$, there is an issue of if it is possible to derive (2.1) from (2.2) (or (2.4) from (2.5)). Without prior restrictions, there can be infinitely many different SVAR that yield identical (2.2). To see this we note that premultiplying (2.1) by any nonsingular constant matrix F yields

$$\tilde{A}_0\underset{\sim}{w}_t + \tilde{A}_1\underset{\sim}{w}_{t-1} + \cdots + \tilde{A}_p\underset{\sim}{w}_{t-p} = \underset{\sim}{\tilde{\epsilon}}_t, \tag{2.6}$$

where $\tilde{A}_j = FA_j, \underset{\sim}{\tilde{\epsilon}}_t = F\underset{\sim}{\epsilon}_t$. Equations (2.1) and (2.5) yield identical (2.2) since $\tilde{A}_0^{-1}\tilde{A}_j = A_0^{-1}F^{-1}FA_j = \Pi_j, \underset{\sim}{v}_t = \tilde{A}_0^{-1}\underset{\sim}{\tilde{\epsilon}}_t = A_0^{-1}F^{-1}F\underset{\sim}{\epsilon}_t = A_0^{-1}\underset{\sim}{\epsilon}_t$. In other words, (2.1) and (2.5) are *observationally equivalent*.

An equation in (2.1) is identified if and only if the g-th row of *admissible transformation* matrix $F = (\underset{\sim}{f}_g')$ takes the form that apart from the gth element being a nonzero constant, the rest are all zeros, i.e., $\underset{\sim}{f}_g' = (0, \ldots, 0, f_{gg}, 0, \ldots, 0)$ (e.g. Hsiao [23]). The transformation matrix F is *admissible* if and only if (2.1) and (2.6) satisfy the same prior restrictions. Suppose that the g-th equation of (2.1) satisfies the prior restrictions $\underset{\sim}{a}_g'\Phi_g = \underset{\sim}{0}'$, where $\underset{\sim}{a}_g'$ denotes the g-th row of A and Φ_g denotes a $(p+1)m \times R_g$ matrix with known elements. Let $\Phi_g^* = M^{-1}\Phi_g$, the existence of prior restrictions $\underset{\sim}{a}_g'\Phi_g = \underset{\sim}{0}'$ is equivalent to the existence of prior restrictions $\underset{\sim}{a}_g^{*'}\Phi_g^* = \underset{\sim}{0}'$, where $\underset{\sim}{a}_g^{*'}$ is the g-th row of A^*. It is shown by Hsiao [26] that

[3]Since $\Pi_p^* = A_0^{-1}A_p^*$, A4 implies that (a) $\Pi_p^* = \underset{\sim}{\alpha}^*\underset{\sim}{\beta}^{*'}$, where $\underset{\sim}{\alpha}^*$ and $\underset{\sim}{\beta}^*$ are $m \times r$ matrices of full column rank $r, 0 \leq r \leq m-1$, and (b) $\underset{\sim}{\alpha}_\perp^{*'} J^* \underset{\sim}{\beta}_\perp^*$ is nonsingular, where $J^* = \sum_{j=0}^{p-1} \Pi_j^*$.

Theorem 2.1. *Suppose that the g-th equation of (2.1) is subject to the prior restrictions* $\underset{\sim}{a}_g'\Phi_g = \underset{\sim}{0}'$. *A necessary and sufficient condition for the identification of the g-th equation of (2.1) or (2.4) is that*

$$\operatorname{rank}(A\Phi_g) = m - 1, \tag{2.7}$$

or

$$\operatorname{rank}(A^*\Phi_g^*) = m - 1. \tag{2.8}$$

Let $\underset{\sim}{w}_t' = (\underset{\sim}{y}_t', \underset{\sim}{x}_t')$, where $\underset{\sim}{y}_t'$ and $\underset{\sim}{x}_t'$ are $1 \times G$ and $1 \times K$, respectively, and $G + K = m$. Let

$$A(L) = \begin{bmatrix} A_{11}(L)\ A_{12}(L) \\ A_{21}(L)\ A_{22}(L) \end{bmatrix},$$

and $\underset{\sim}{\epsilon}_t' = (\underset{\sim}{\epsilon}_{1t}', \underset{\sim}{\epsilon}_{2t}')$ be the conformable partitions. Cowles Commission decomposition of $\underset{\sim}{w}_t$ into joint dependent variable variables $\underset{\sim}{y}_t$ and exogenous variables $\underset{\sim}{x}_t$ is equivalent to imposing the prior restrictions (Zellner and Palm [60]),

$$A_{21}(L) \equiv \underset{\sim}{0} \quad \text{and} \quad E\underset{\sim}{\epsilon}_{1t}\underset{\sim}{\epsilon}_{2t}' = \underset{\sim}{0}. \tag{2.9}$$

The prior restrictions (2.9) restrict the *admissible transformation matrix* F to be block diagonal (e.g. Hsiao [23]). Therefore,

Corollary 2.1. *Under (2.9) and* $\underset{\sim}{a}_g'\Phi_g = \underset{\sim}{0}'$, *a necessary and sufficient condition for the identification of the g-th equation for* $g \leq G$ *is*

$$\operatorname{rank}[(A_{11} \quad A_{12})\Phi_g] = G - 1, \tag{2.10}$$

where A_{11} *and* A_{12} *are conformable partitions of A.*

The identification condition (2.7) or (2.8) does not require any prior knowledge of the direction of nonstationarity or the rank of cointegration. As a matter of fact many macroeconometric models are identified without any prior knowledge of location of unit roots or rank of cointegration, (e.g. the Klein [34] interwar model and the large scale Wharton quarterly model (Klein and Evans [35]). Of course, if such information is available, it can improve the efficiency of system estimators and simplify the issues of inference considerably (e.g. King, Plosser, Stock and Watson [33]).

3. Inference in VAR (or reduced form)

Consider the g-th equation of (2.2),

$$\underset{\sim}{w}_g = X\underset{\sim}{\pi}_g + \underset{\sim}{v}_g, \tag{3.1}$$

where $\underset{\sim}{w}_g$ is the $T \times 1$ vector of the g-th element of $\underset{\sim}{w}_t, w_{gt}, X = (W_{-1}, \ldots, W_{-p})$, is the $T \times mp$ vector of $\underset{\sim}{w}_{t-1}, \ldots, \underset{\sim}{w}_{t-p}, \underset{\sim}{\pi}_g$ is the corresponding vector of coefficients, and $\underset{\sim}{v}_g$ is the $T \times 1$ vector of the g-th element of $\underset{\sim}{v}_t, v_{gt}$.

Rewrite (3.1) in terms of linearly independent $I(0)$ and full rank $I(1)$ regressors X_1^* and X_2^*, respectively, by postmultiplying a nonsingular transformation matrix

M_x to X,[4] we have

$$\begin{aligned} \underset{\sim}{w}_g &= XM_xM_x^{-1}\underset{\sim}{\pi}_g + \underset{\sim}{v}_g \\ &= X^*\underset{\sim}{\pi}_g^* + \underset{\sim}{v}_g \\ &= (X_1^*, X_2^*)\begin{pmatrix} \underset{\sim}{\pi}_{g1}^* \\ \underset{\sim}{\pi}_{g2}^* \end{pmatrix} + \underset{\sim}{v}_g, \end{aligned} \tag{3.2}$$

where $\underset{\sim}{\pi}_g^* = M_x^{-1}\underset{\sim}{\pi}_g = (\underset{\sim}{\pi}_{g1}^{*'}, \underset{\sim}{\pi}_{g2}^{*'})'$. The least squares estimator of (3.1) is equal to M_x times the least squares estimator of (3.2),

$$\begin{aligned} \hat{\underset{\sim}{\pi}} &= (X'X)^{-1}(X'\underset{\sim}{w}_g) \\ &= M_x(X^{*'}X^*)^{-1}X^{*'}\underset{\sim}{w}_g \\ &= M_x[\underset{\sim}{\pi}_g^* + (X^{*'}X^*)^{-1}X^{*'}\underset{\sim}{v}_g]. \end{aligned} \tag{3.3}$$

The statistical properties of (3.3) can be derived by making use of the fundamental functional central limit theorems proved by Chan and Wei [7], Phillips and Durlauf [46], etc.:

Theorem 3.1. *Let $\underset{\sim}{\eta}_t$ be an $m \times 1$ vector of random variables with $E(\underset{\sim}{\eta}_t \mid \underset{\sim}{\eta}_{t-1}, \ldots,) = \underset{\sim}{0}$, $E(\underset{\sim}{\eta}_t\underset{\sim}{\eta}_t' \mid \underset{\sim}{\eta}_{t-1}, \ldots,) = I_m$, and bounded fourth moments. Let $F(L) = \sum_{j=0}^{\infty} F_jL^j$ and $G(L) = \sum_{j=0}^{\infty} G_jL^j$ with $\sum_{j=0}^{\infty} j \mid F_j \mid < \infty$ and $\sum_{j=0}^{\infty} j \mid G_j \mid < \infty$. Let $\underset{\sim}{\xi}_t = \sum_{s=1}^{t} \underset{\sim}{\eta}_s$, and let $B(r)$ denote an $m \times 1$ dimensional Brownian motion process.*

Then

(a) $T^{-1/2}\sum_{t=1}^{T} F(L)\underset{\sim}{\eta}_t \Longrightarrow N(0, F(1)F(1)')$,
(b) $T^{-1}\sum_{t=1}^{T} \underset{\sim}{\xi}_{t-1}\underset{\sim}{\eta}_t' \Longrightarrow \int B(r)dB(r)'$,
(c) $T^{-1}\sum_{t=1}^{T} \underset{\sim}{\xi}_t[F(L)\underset{\sim}{\eta}_t]' \Longrightarrow F(1)' + \int B(r)dB(r)'F(1)'$,
(d) $T^{-1}\sum_{t=1}^{T}[F(L)\underset{\sim}{\eta}_t][G(L)\underset{\sim}{\eta}_t]' \longrightarrow \sum_{j=0}^{\infty} F_jG_j'$,
(e) $T^{-2}\sum_{t=1}^{T} \underset{\sim}{\xi}_t\underset{\sim}{\xi}_t' \Longrightarrow \int B(r)B(r)'dr$,

where to simplify notation $\int_0^1$ is denoted by $\int$ and $\longrightarrow$ and $\Longrightarrow$ denote convergence in probability and distribution of the associated probability measure, respectively.

Making use of theorem 3.1, it follows that

Theorem 3.2. *Under Assumptions A.1 - A.4, as $T \longrightarrow \infty$,*

$$\sqrt{T}(\hat{\underset{\sim}{\pi}}_{g1}^* - \underset{\sim}{\pi}_{g1}^*) \Longrightarrow N(\underset{\sim}{0}, \sigma_{v_g}^2 M_{x_1x_1}^*), \tag{3.4}$$

$$\begin{aligned} T(\hat{\underset{\sim}{\pi}}_{g2}^* - \underset{\sim}{\pi}_{g2}^*) \Longrightarrow &\left(\int B_{x_2^*}(r)B_{x_2^*}(r)'dr\right)^{-1} \\ &\left(\int B_{x_2^*}(r)dB_{v_g}(r)\right). \end{aligned} \tag{3.5}$$

where $M_{x_1x_1}^* = plim\ \frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{x}_{1t}^*\underset{\sim}{x}_{1t}^{*'}$. Moreover, (3.4) and (3.5) are asymptotically independent.

[4]Such a transformation always exist. However, it does not need to be known *a priori*. The use of (3.2) is to facilitate the derivation of statistical properties of the estimators of (3.1) or (2.1).

The least squares estimator (3.3) is a linear combination of $\hat{\underset{\sim}{\pi}}^*_{g1}$ and $\hat{\underset{\sim}{\pi}}^*_{g2}$. Its limiting distribution is determined by the limiting distribution of the slower rate of $\hat{\underset{\sim}{\pi}}^*_g$ included. Since the limiting distribution of $\hat{\underset{\sim}{\pi}}^*_{g2}$ is nonstandard and involves a matrix unit distribution, the usual Wald test statistic under the null may not be approximated by the chi-square distribution if the null hypothesis involves coefficients in the direction of nonstationarity (e.g. Dolado and Lutkepohl [9], Sims, Stock and Watson [51], Tsay and Tiao [57]). On the other hand, if $\underset{\sim}{w}_t$ is cointegrated and the rank of cointegration is known *a priori*, Ahn and Reinsel [1] and Johansen [31, 32] using the reduced rank framework proposed by Anderson [3] have shown that the coefficients of cointegration vectors are asymptotically mixed normal, hence there will be no inference problem. The Wald test statistics constructed from the reduced rank regression will again be asymptotically chi-square distributed. This is because imposing the reduced rank condition is equivalent to avoid estimating the unit roots in the system.

Unfortunately, as discussed in section 2, prior information on the rank of cointegration or direction of nonstationarity is usually lacking. One way to deal with it is to pretest the data for the presence of cointegration and the rank of cointegration, then apply the reduced rank regression of Ahn and Reinsel [1] or Johansen [31, 32]. However, statistic tests for the rank of cointegration have very poor finite sample performance (e.g. Stock [53]). The first stage unit root test and second stage cointegration test can induce substantial size distortion. For instance, Elliott and Stock [10] consider a bivariate problem in which there is uncertainty about whether the regressor has a unit root. In their Monte Carlo simulation they find that unit root pretests can induce substantial size distortions in the second-stage test. If the innovations of the regressors and the second-stage regression error are correlated, the first-stage Dickey-Fuller [8] t-statistic and the second-stage t-statistic will be dependent so the size of the second stage in this two-stage procedure cannot be controlled, even asymptotically. Many other Monte Carlo studies also show that serious size and power distortions arise and the number of linearly independent cointegrating vectors tend to be overestimated as the dimension of the system increases relative to the time dimension (e.g. Ho and Sorensen [18], Gonzalo and Pitarakis [12]).

Another way is to correct the miscentering and skewness of the limiting distribution of the least squares estimator due to the "endogeneities" of the predetermined integrated regressors (e.g. Park [42], Phillips [44], Phillips and Hansen [47], Robinson and Hualde [49]). However, since the rank of cointegration and direction of nonstationarity are unknown, Phillips [45] proposes to deal with potential endogeneities by making a correction of the least squares regression formula that adjusts for whatever endogeneities there may be in the predetermined variables that is due to their nonstationarity by transforming the dependent variables $\underset{\sim}{w}_t$ into

$$\underset{\sim}{w}^+_t = \underset{\sim}{w}_t - \Omega_{v\bigtriangledown w}\Omega^-_{\bigtriangledown w\bigtriangledown w} \bigtriangledown \underset{\sim}{w}_t, \tag{3.6}$$

where $\Omega_{\bigtriangledown w\bigtriangledown w} = \sum_{j=-\infty}^{\infty} E(\bigtriangledown \underset{\sim}{w}_t \bigtriangledown \underset{\sim}{w}'_{t-j})$, $\Omega_{v\bigtriangledown w} = \sum_{j=-\infty}^{\infty} E(\underset{\sim}{v}_t \bigtriangledown \underset{\sim}{w}'_{t-j})$ and $\Omega^-_{\bigtriangledown w\bigtriangledown w}$ denotes the Moore-Penrose generalized inverse.[5] Using $\underset{\sim}{w}^+_t$ in place of $\underset{\sim}{w}_t$ in (2.2) is equivalent to modifying the error term from $\underset{\sim}{v}_t$ to $\underset{\sim}{v}_t - \Omega_{\bigtriangledown w}\Omega^-_{\bigtriangledown w} \bigtriangledown \underset{\sim}{w}_t$, which now becomes serially correlated because $\bigtriangledown \underset{\sim}{w}_t$ is serially correlated. To correct for this order (1/T) serial correlation bias term, Phillips [45] suggests further adding $(X'X)^{-1}(\underset{\sim}{0}, T\Delta^{+'}_{v\bigtriangledown w})$ to the least squares regression estimator of $\underset{\sim}{w}^+_t$ on $\bigtriangledown \underset{\sim}{w}_{t-1}, \ldots, \bigtriangledown \underset{\sim}{w}_{t-p+1}, \underset{\sim}{w}_{t-p}$, where $\Delta^+_{v\bigtriangledown w} = \Omega_{v\bigtriangledown w}\Omega^-_{\bigtriangledown w\bigtriangledown w}\Delta_{\bigtriangledown w\bigtriangledown w}$, and Δ_{uv}

[5] If $\underset{\sim}{w}_t$ are cointegrated, $\Omega_{\bigtriangledown w\bigtriangledown w}$ does not have full rank.

denotes the one-sided long-run covariances of two sets of $I(0)$ variables (u_t, v_t), $\Delta_{uv} = \sum_{j=0}^{\infty} \Gamma_{uv}(j)$ where $\Gamma_{uv}(j) = E\underset{\sim}{u}_t\underset{\sim}{v}_{t-j}$.[6] Consistent estimates of Ω_{uv} or Δ_{uv} can be obtained by using Kernel method (e.g. Hannan [15], Priestley [48]).

$$\hat{\Omega}_{uv} = \sum_{j=-T+1}^{T-1} h(j/K)\hat{\Gamma}_{uv}(j), \tag{3.7}$$

$$\hat{\Delta}_{uv} = \sum_{j=0}^{T-1} h(j/K)\hat{\Gamma}_{uv}(j), \tag{3.8}$$

where $\hat{\Gamma}_{uv}(j)$ is a consistent sample covariance estimator of $\Gamma_{uv}(j)$, and $h(\cdot)$ is a kernel function and K is a lag truncation or bandwidth parameter. Assuming that

Assumption 3.1. The kernel function $h(\cdot) : R \longrightarrow [-1, 1]$ is a twice continuously differentiable even function with:

(a) $h(0) = 1, h'(0) = 0, h''(0) \neq 0$; and either
(b) $h(x) = 0, \mid x \mid \geq 1$, with $\lim_{|x| \longrightarrow 1} \frac{h(x)}{(1-|x|)^2} =$ constant, or
(b') $h(x) = O((1-x)^2)$, as $\mid x \mid \longrightarrow 1$.

Assumption 3.2. The bandwidth parameter K in the kernel estimates (3.7) and (3.8) has an expansion rate $K \sim c_T T^k$ for some $k \in (1/4, 2/3)$ and for some slowly varying function c_T and thus $K/T^{2/3} + T^{1/4}/K \longrightarrow 0$ and $K^{4/T} \longrightarrow \infty$ as $T \longrightarrow \infty$.

Phillips [45] shows that the modified least squares estimates are either asymptotically normally distributed or mixed normal. However, because the direction of nonstationarity is unknown, the conditional covariance matrix cannot be derived. Therefore, if the test statistic involves some of the coefficients of nonstationary variables, the limiting distribution becomes a mixture of chi-squares variates with the weights between 0 and 1. In other words, if tests based on chi-square distribution reject the null with significance level α, then the test rejects the null with significance level less than α. In other words, tests based on chi-square distribution provides a conservative test.

Toda and Yamamoto [55] have suggested a lag-order augmented approach to circumscribe the issue of non-standard distributions associated with integrated regressors by overfitting a VAR with additional $d_{\max}$ lags where $d_{\max}$ denotes the maximum order of integration suspected. In our case, $d_{\max} = 1$. In other words, instead of estimating (2.2), we estimate

$$\underset{\sim}{w}_t = \Pi_1\underset{\sim}{w}_{t-1} + \cdots + \Pi_p\underset{\sim}{w}_{t-p} + \Pi_{p+1}\underset{\sim}{w}_{t-p-1} + \underset{\sim}{v}_t, \tag{3.9}$$

Since we know *a priori*, $\Pi_{p+1} \equiv 0$, we are only interested in the estimates of $\Pi_j, j = 1, \ldots, p$. The limiting distributions of the least squares estimates of (3.9) can be derived from the limiting distributions of the least squares estimates of (the error-correction form),

$$\underset{\sim}{w}_t = \Pi_1^* \bigtriangledown \underset{\sim}{w}_{t-1} + \cdots + \Pi_p^* \bigtriangledown \underset{\sim}{w}_{t-p} + \Pi_{p+1}^*\underset{\sim}{w}_{t-p-1} + \underset{\sim}{v}_t, \tag{3.10}$$

because $\Pi_j^* = \sum_{i=1}^{j} \Pi_i, j = 1, \ldots, p+1$ or $\Pi_j = \Pi_j^* - \Pi_{j-1}^*$ where $\Pi_0^* \equiv \underset{\sim}{0}$. Since $\Pi_j^*, j = 1, \ldots, p$ are coefficients of stationary regressors, Theorem 3.2 shows that the

[6] Under A.3, $\Delta_{v \bigtriangledown w} = \underset{\sim}{0}$.

least squares estimates of $\Pi_j^*, j = 1, \ldots, p$ converge to the true values at the speed of $\sqrt{T}$ and are asymptotically normally distributed. Only the least squares estimates of Π_{p+1}^* may be T-convergent and have non-normal limiting distributions. However, since we know *a priori* that $\Pi_{p+1} = \underset{\sim}{0}$, our interest is only in $\Pi_j, j = 1, \ldots, p$. The least squares regression of (3.9) yields $\hat{\Pi}_j = \hat{\Pi}_j^* - \hat{\Pi}_{j-1}^*, j = 1, \ldots, p$, therefore, they are asymptotically normally distributed. Wald test statistics of the null hypothesis constructed from regression estimates of (3.9) will again be asymptotically chi-square distributed.

Phillips [45] modified estimator maintains the T-convergence part of the coefficients associated with full rank integrated regressors. The Toda-Yamamoto [55] lag order augmented estimator is only $\sqrt{T}$-convergent. So Phillips [45] modified estimator is likely to be asymptotically more efficient. However, computationally, the Phillips modified estimator is much more complicated than the lag order augmented estimator. Moreover, test statistics constructed from the modified estimators can only give the bounds of the size of the test because the conditional variance is unknown, while test statistics constructed from the lag order augmented estimator asymptotically yield the exact size.

4. Least squares and two stage least squares estimation of SVAR

For ease of exposition, we assume that prior information is in the form of excluding certain variables, both current and lagged, from an equation. Let the g-th equation of (2.1) be written as

$$\underset{\sim}{w}_g = Z_g \underset{\sim}{\delta}_g + \underset{\sim}{\epsilon}_g, \tag{4.1}$$

where $\underset{\sim}{w}_g$ and $\underset{\sim}{\epsilon}_g$ denote the $T \times 1$ vectors of $(w_{g1}, \ldots, w_{gT})'$ and $(\epsilon_{g1}, \ldots, \epsilon_{gT})'$, respectively, and Z_g denotes the $T \times [(p+1)g_\Delta - 1]$ dimensional matrix of g_Δ included current and lagged variables of $\underset{\sim}{w}_t$.

The least squares estimator of (4.1) is given by

$$\hat{\underset{\sim}{\delta}}_{g,\ell s} = (Z_g' Z_g)^{-1} Z_g' \underset{\sim}{w}_g \tag{4.2}$$

Phillips and Durlauf [46] and Stock [52] have shown that the least squares estimator with integrated regressors is consistent even when the regressors and the errors are correlated. However, the basic assumption underlying their result is that the regressors are not cointegrated. In a dynamic framework even though $\underset{\sim}{w}_{t-j}$ are $I(1)$, the current and lagged variables are trivially cointegrated. It was shown in [21] when contemporaneous joint dependent variables also appear as explanatory variables in (4.1), applying least squares method to (4.1) does not yield consistent estimator for $\underset{\sim}{\delta}_g$. To see this, let M_g be the nonsingular transformation matrix that transforms Z_g into $Z_g^* = Z_g M_g = (Z_{g1}^*, Z_{g2}^*)$, where Z_{g1}^* denotes the ℓ_g-dimensional linearly independent $I(0)$ variables and Z_{g2}^* denotes the T observations of b_g full rank $I(1)$ variables,[7] then

$$\begin{aligned} \underset{\sim}{w}_g &= Z_g M_g M_g^{-1} \underset{\sim}{\delta}_g + \underset{\sim}{\epsilon}_g \\ &= Z_g^* \underset{\sim}{\delta}_g^* + \underset{\sim}{\epsilon}_g \end{aligned} \tag{4.3}$$

where $\underset{\sim}{\delta}_g^* = M_g^{-1} \underset{\sim}{\delta}_g = (\underset{\sim}{\delta}_{g1}^{*'}, \underset{\sim}{\delta}_{g2}^{*'})'$ with $\underset{\sim}{\delta}_{g1}^*$ and $\underset{\sim}{\delta}_{g2}^*$ denoting the $\ell_g \times 1$ and $b_g \times 1$ vector, respectively. Such transformation always exists. For instance, if no cointegrating relation exists among the included $\underset{\sim}{w}_t$, say $\tilde{\underset{\sim}{w}}_{gt}$, then b_g equals the dimension

[7] By full rank I(1) variables we mean that there is no cointegrating relation among Z_{g2}^*.

of included joint dependent variables, g_Δ, and Z^*_{g1} consists of the first differenced current and $p-1$ lagged included variables, Z^*_{g2} is simply the $T \times b_g$ (or $T \times g_\Delta$) included $\tilde{\underset{\sim}{w}}_{gt}$ lagged by p periods, $\tilde{\underset{\sim}{w}}_{g,t-p}$. On the other hand, if there exists $g_\Delta - b_g$ linearly independent cointegrating relations among the g_Δ included variables, $\tilde{\underset{\sim}{w}}_{gt}$, then Z^*_{g1} consists of the current and $p-1$ lagged $\bigtriangledown\tilde{\underset{\sim}{w}}_{gt}$ and $\tilde{W}_{g,-p}\underset{\sim}{d}_g$ cointegrating relations, where $\tilde{W}_{g,-p}$ is $T \times g_\Delta$ matrix of included $\tilde{\underset{\sim}{w}}_{g,t-p}$, $\underset{\sim}{d}_g$ is $g_\Delta \times (g_\Delta - b_g)$ of constants, and Z^*_{g2} consists of the T observed b_g full rank $I(1)$ variables $\tilde{W}_{g2,-p}$.

The least squares estimator (4.2) can be written as $\hat{\underset{\sim}{\delta}}_{g,\ell s} = M_g \hat{\underset{\sim}{\delta}}^*_{g,\ell s}$, where $\hat{\underset{\sim}{\delta}}^*_{g,\ell s}$ denotes the least squares estimator of (4.3). Using Theorem 3.1, one can show that $\frac{1}{T} Z^{*'}_{g1} Z^*_{g1} \longrightarrow M^*_{z_{g1}z_{g1}}, T^{-2/3} Z^{*'}_{g1} Z^*_{g2} \longrightarrow \underset{\sim}{0}, \frac{1}{T^2} Z^{*'}_{g2} Z^*_{g2} \Longrightarrow M^*_{z_{g2}z_{g2}}, \frac{1}{T^2} Z^{*'}_{g2} \underset{\sim}{\epsilon}_g \longrightarrow \underset{\sim}{0}$, $\frac{1}{T} Z^*_{g1} \underset{\sim}{\epsilon}_g \longrightarrow \underset{\sim}{b}$, where $\underset{\sim}{b} = [E(\underset{\sim}{\epsilon}_{gt} \tilde{\underset{\sim}{w}}'_{gt}), \underset{\sim}{0}']' = [(A_0^{-1} \sum_{\epsilon\epsilon,g})'_g, \underset{\sim}{0}']'$, $\sum_{\epsilon\epsilon,g}$ is the g-th column of $\sum_{\epsilon\epsilon}$ and $(A_0^{-1} \sum_{\epsilon\epsilon,g})_g$ is the $(g_\Delta - 1) \times 1$ subvector of $A_0^{-1} \sum_{\epsilon\epsilon,g}$ that corresponds to the $g_\Delta - 1$ included variables $\tilde{\underset{\sim}{w}}_{gt}$ in the g-th equation, and $M^*_{z_{g1}z_{g1}}$ and $M^*_{z_{g2}z_{g2}}$ are nonsingular. It follows that

$$\hat{\underset{\sim}{\delta}}^*_{g,\ell s} = \begin{bmatrix} \hat{\underset{\sim}{\delta}}^*_{g1,\ell s} \\ \hat{\underset{\sim}{\delta}}^*_{g2,\ell s} \end{bmatrix} \longrightarrow \begin{bmatrix} \underset{\sim}{\delta}^*_{g1} \\ \underset{\sim}{\delta}^*_{g2} \end{bmatrix} + \begin{bmatrix} \underset{\sim}{b} \\ \underset{\sim}{0} \end{bmatrix}. \tag{4.4}$$

Although the coefficients of Z^*_{g2} can be consistently estimated, the coefficients of Z^*_{g1} cannot. Since $\hat{\underset{\sim}{\delta}}_{g,\ell s}$ is a linear combination of $\hat{\underset{\sim}{\delta}}^*_{g1,\ell s}$ and $\hat{\underset{\sim}{\delta}}^*_{g2,\ell s}, \hat{\underset{\sim}{\delta}}_{g,\ell s}$ is inconsistent.

When the errors and regressors are correlated, a standard procedure is to use instrumental variable method. Using lagged variables as instruments, the two stage least squares estimator of $\underset{\sim}{\delta}_g$ is given by

$$\hat{\underset{\sim}{\delta}}_{g,2SLS} = [Z'_g X (X'X)^{-1} X' Z_g]^{-1} [Z'_g X (X'X)^{-1} Z'_g \underset{\sim}{w}_g], \tag{4.5}$$

where $X = (W_{-1}, W_{-2}, \ldots, W_{-p})$ and W_{-j} denotes the $T \times m$ matrix representation of $\underset{\sim}{w}_{t-j}$. Transforming X into linearly independent $I(0)$ and full rank $I(1)$ processes, X^*_1 and X^*_2, respectively, by $M_x, XM_x = [X^*_1, X^*_2]$, the 2SLS estimator (4.5) is equal to $M_g \hat{\underset{\sim}{\delta}}^*_{g,2SLS}$, where

$$\hat{\underset{\sim}{\delta}}^*_{g,2SLS} = [Z^{*\prime}_g X^* (X^{*\prime} X^*)^{-1} X^{*\prime} Z^*_g]^{-1} [Z^{*\prime}_g X^* (X^{*\prime} X^*)^{-1} X^{*\prime} \underset{\sim}{w}_g] \tag{4.6}$$

Since $\frac{1}{T^2} Z^{*\prime}_{g1} X^*_2 \longrightarrow \underset{\sim}{0}$, $\frac{1}{T} Z^{*\prime}_{g2} X^*_1 \Longrightarrow M^*_{z_{g2}x_1}$, $\frac{1}{T^2} Z^{*\prime}_{g2} X^*_1 \longrightarrow 0$, $\frac{1}{T} X^{*\prime}_1 X^*_1 \longrightarrow M^*_{x_1x_1}$, $\frac{1}{T} X^{*\prime}_1 X^*_2 \Longrightarrow M^*_{x_1x_2}$ $\frac{1}{T^2} X^{*\prime}_1 X^*_2 \longrightarrow 0$, $\frac{1}{T^2} X^{*\prime}_2 X^*_2 \Longrightarrow M^*_{x_2x_2}$ $\frac{1}{T} X^{*\prime}_1 \underset{\sim}{\epsilon}_g \longrightarrow \underset{\sim}{0}$, and $\frac{1}{T^2} X^{*\prime}_2 \underset{\sim}{\epsilon}_g \longrightarrow \underset{\sim}{0}$, and $M^*_{x_2x_2}$ are nonsingular, it follows that $\hat{\underset{\sim}{\delta}}^*_{g,2SLS}$ converges to $\underset{\sim}{\delta}^*_g$. Hence the 2SLS estimator of $\underset{\sim}{\delta}_g$ is consistent.

Let $H_g = \begin{bmatrix} T^{-\frac{1}{2}} I_{\ell_g} & \underset{\sim}{0} \\ \underset{\sim}{0} & T^{-1} I_{b_g} \end{bmatrix}$ and $H_x = \begin{bmatrix} T^{-\frac{1}{2}} I_{\ell^*} & \underset{\sim}{0} \\ \underset{\sim}{0} & T^{-1} I_{b^*} \end{bmatrix}$, where ℓ^* and b^* are the column dimensions of X^*_1 and X^*_2 respectively. Under assumptions A.1 - A.4, as $T \longrightarrow \infty$,

$$\begin{aligned} H_g^{-1} (\hat{\underset{\sim}{\delta}}^*_{g,2SLS} - \underset{\sim}{\delta}^*_g) &= \begin{bmatrix} \sqrt{T} (\hat{\underset{\sim}{\delta}}^*_{g1,2SLS} - \underset{\sim}{\delta}^*_{g1}) \\ T (\hat{\underset{\sim}{\delta}}^*_{g2,2SLS} - \underset{\sim}{\delta}^*_{g2}) \end{bmatrix} \\ &\Longrightarrow \begin{bmatrix} (M^*_{z_{g1}x_1} M^{*-1}_{x_1x_1} M^*_{x_1z_{g1}})^{-1} (M^*_{z_{g1}x_1} M^{*-1}_{x_1x_1} \cdot T^{-1/2} X^{*\prime}_1 \underset{\sim}{\epsilon}_g) \\ (M^*_{z_{g2}x_2} M^{*-1}_{x_2x_2} M^*_{x_2z_{g2}})^{-1} (M^*_{z_{g2}x_2} M^{*-1}_{x_2x_2} \cdot T^{-1} X^{*\prime}_2 \underset{\sim}{\epsilon}_g) \end{bmatrix}. \end{aligned} \tag{4.7}$$

By theorem 3.1, we have

$$\frac{1}{\sqrt{T}} X^{*\prime}_1 \underset{\sim}{\epsilon}_g \Longrightarrow N(\underset{\sim}{0}, \sigma^2_g M^*_{x_1x_1}). \tag{4.8}$$

and

$$\frac{1}{T}X_2^{*\prime}\underset{\sim}{\epsilon}_g \Longrightarrow \int B_{x_2^*}dB_{\epsilon_g}, \tag{4.9}$$

where B_{ϵ_g} denotes the Brownian motion of ϵ_{gt} with variance σ_g^2, $B_{x_2^*}$ denotes a $b^* \times 1$ vector Brownian motion of $\bigtriangledown \underset{\sim}{x}_{2t}^*$ with covariance matrix $\Omega_{\bigtriangledown x_2^* \bigtriangledown x_2^*}$ where $\Omega_{\bigtriangledown x_2^* \bigtriangledown x_2^*}$ is the long-run covariance matrix of $\bigtriangledown \underset{\sim}{x}_{2t}^*$. The Brownian motion $B_{x_2^*}^*$ and B_{ϵ_g} are not independent because ϵ_{gt} and $\underset{\sim}{v}_t$ are contemporaneously correlated. Following Phillips [44], we can decompose the right hand side of (4.9) into two terms as

$$\int B_{x_2^*}dB_{\epsilon_g \cdot x_2^*} + \int B_{x_2^*}\Omega_{\epsilon_g \bigtriangledown x_2^*}\Omega^{-1}_{\bigtriangledown x_2^* \bigtriangledown x_2^*}dB_{x_2^*}, \tag{4.10}$$

where $B_{\epsilon_g \cdot x_2^*} = B_{\epsilon_g} - \Omega_{\epsilon_g \bigtriangledown x_2^*}\Omega^{-1}_{\bigtriangledown x_2^* \bigtriangledown x_2^*}B_{x_2^*} \equiv BM(\sigma^2_{g \cdot \bigtriangledown x_2^*})$ with $\sigma^2_{g \cdot \bigtriangledown x_2^*} = \sigma_g^2 - \Omega_{\epsilon_g \bigtriangledown x_2^*}\ \Omega^{-1}_{\bigtriangledown x_2^* \bigtriangledown x_2^*}\Omega_{\bigtriangledown x_2^* \epsilon_g}$, and $\Omega_{\epsilon_g \bigtriangledown x_2^*}$ denotes the long-run covariance between $\underset{\sim}{\epsilon}_g$ and $\bigtriangledown \underset{\sim}{x}_2^*$. The first term of (4.10) is a mixed normal. The second term involves a matrix unit root distribution that arises from using lagged $\underset{\sim}{w}$ as instruments when $\underset{\sim}{w}$ is I(1) and the contemporaneous correlation between $\underset{\sim}{\epsilon}_{gt}$ and $\underset{\sim}{w}_t$ is nonzero. The "long-run endogeneity" of the nonstationary instruments X_2^* leads to a skewness of the limiting distribution of $\hat{\underset{\sim}{\delta}}^*_{g,2SLS}$ and its dependence on nuisance parameters that are impossible to eliminate by the 2SLS. Therefore,

Theorem 4.1. *Under A.1 - A.4 the 2SLS estimator of $\underset{\sim}{\delta}_g^*$ is consistent and*

$$\sqrt{T}(\hat{\underset{\sim}{\delta}}^*_{g1,2SLS} - \underset{\sim}{\delta}^*_{g1}) \Longrightarrow N(\underset{\sim}{0}, \sigma_g^2(M^*_{z_{g1}x_1}M^{*-1}_{x_1x_1}M^*_{x_1z_{g1}})^{-1}), \tag{4.11}$$

$$\begin{aligned} T(\hat{\underset{\sim}{\delta}}^*_{g2,2SLS} - \underset{\sim}{\delta}^*_{g2}) \Longrightarrow & \left\{\int B_{z^*_{g2}}B'_{x_2^*}dr(\int B_{x_2^*}B'_{x_2^*}dr)^{-1}\int B_{x_2^*}B'_{z^*_{g2}}dr\right\}^{-1} \\ & \left\{\int B_{z^*_{g2}}B'_{x_2^*}dr(\int B_{x_2^*}B'_{x_2^*}dr)^{-1}\right. \\ & \left.\times\left[\int B_{x_2^*}dB_{\epsilon_g \cdot x_2^*} + \int B_{x_2^*}\Omega_{\epsilon_g \bigtriangledown x_2^*}\Omega^{-1}_{\bigtriangledown x_2^* \bigtriangledown x_2^*}dB_{x_2^*}\right]\right\}, \end{aligned} \tag{4.12}$$

where $B_{z^*_{g2}}$ denotes a $b_g \times 1$ vector Brownian motion of $\bigtriangledown \underset{\sim}{z}^*_{g2,t}$ which appears in the g-th equation. The distributions of (4.11) and (4.12) are asymptotically independent.

Theorem 4.1 suggests that inference about the null hypothesis $P\underset{\sim}{\delta}_g = \underset{\sim}{c}$ can be tricky, where P and $\underset{\sim}{c}$ are known matrix and vector of proper dimensions. If $\sqrt{T}P(\hat{\underset{\sim}{\delta}}_{g,2SLS} - \underset{\sim}{\delta}_g)$ has a nonsingular covariance matrix, the limiting distribution of $P\hat{\underset{\sim}{\delta}}_g$ is determined by the limiting distribution of $\hat{\underset{\sim}{\delta}}^*_{g1}$, hence the Wald test statistic

$$(\hat{\underset{\sim}{\delta}}_{g,2SLS} - \underset{\sim}{\delta}_g)'P' \text{ Cov } (P\hat{\underset{\sim}{\delta}}_{g,2SLS})^{-1}P(\hat{\underset{\sim}{\delta}}_{g,2SLS} - \underset{\sim}{\delta}_g) \tag{4.13}$$

under the null will be asymptotically chi-square distributed. On the other hand, if $\sqrt{T}P(\hat{\underset{\sim}{\delta}}_{g,2SLS} - \underset{\sim}{\delta}_g)$ has a singular covariance matrix, it means that there exists a nonsingular matrix L such that

$$LP\underset{\sim}{\delta}_g = LP^*\underset{\sim}{\delta}^*_g = \begin{bmatrix} \tilde{P}_{11} & \tilde{P}_{12} \\ \underset{\sim}{0} & \tilde{P}_{22} \end{bmatrix}\begin{bmatrix} \underset{\sim}{\delta}^*_{g1} \\ \underset{\sim}{\delta}^*_{g2} \end{bmatrix} \tag{4.14}$$

with nonzero $\tilde{P}_{22}$. Then

$$
\begin{aligned}
&(P\hat{\underset{\sim}{\delta}}_{g,2SLS} - \underset{\sim}{c})' \text{ Cov } (P\hat{\underset{\sim}{\delta}}_{g,2SLS})^{-1}(P\hat{\underset{\sim}{\delta}}_{g,2SLS} - \underset{\sim}{c}) \\
&\quad = \left\{ \begin{bmatrix} \tilde{P}_{11} & \tilde{P}_{12} \\ \underset{\sim}{0} & \tilde{P}_{22} \end{bmatrix} \begin{bmatrix} \hat{\underset{\sim}{\delta}}^{*}_{g1,2SLS} \\ \hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS} \end{bmatrix} - L\underset{\sim}{c} \right\}' \text{ Cov } (LP\hat{\underset{\sim}{\delta}}_{g,2SLS})^{-1} \\
&\qquad \times \left\{ \begin{bmatrix} \tilde{P}_{11} & \tilde{P}_{12} \\ \underset{\sim}{0} & \tilde{P}_{22} \end{bmatrix} \begin{bmatrix} \hat{\underset{\sim}{\delta}}^{*}_{g1,2SLS} \\ \hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS} \end{bmatrix} - L\underset{\sim}{c} \right\} \\
&\quad \Longrightarrow T(\tilde{P}_{11}\hat{\underset{\sim}{\delta}}^{*}_{g1,2SLS} + \tilde{P}_{12}\hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS} - \tilde{\underset{\sim}{c}}_1)' \text{ Cov } (\sqrt{T}\tilde{P}_{11}\hat{\underset{\sim}{\delta}}^{*}_{g1,2SLS})^{-1} \\
&\qquad \times (\tilde{P}_{11}\hat{\underset{\sim}{\delta}}^{*}_{g1,2SLS} + \tilde{P}_{12}\hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS} - \tilde{\underset{\sim}{c}}_1) \\
&\qquad + T^2(\tilde{P}_{22}\hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS} - \tilde{\underset{\sim}{c}}_2)' \text{ Cov } (T\tilde{P}_{22}\hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS})^{-1}(\tilde{P}_{22}\hat{\underset{\sim}{\delta}}^{*}_{g2,2SLS} - \tilde{\underset{\sim}{c}}_2),
\end{aligned} \tag{4.15}
$$

where $L\underset{\sim}{c} = (\tilde{\underset{\sim}{c}}_1', \tilde{\underset{\sim}{c}}_2')'$. The first term on the right hand side of (4.15) is asymptotically chi-square distributed. The second term, according to Theorem 3.1 has a nonstandard distribution. Hence (4.15) is not asymptotically chi-square distributed.

If there exists prior information that satisfies (2.9) and $\underset{\sim}{w}_1$ and $\underset{\sim}{w}_2$ are cointegrated with x_2^* contained in $\underset{\sim}{w}_2$, it was shown by Hsiao [22] that the 2SLS converges to a mixed normal distribution. Then the Wald test statistic (4.13) can again be approximated by a chi-square distribution. When variables cannot be dichotomized into "endogenous" and "exogenous", if we do not know the direction of nonstationarity, nor the rank of cointegration, we will not be able to know *a priori* if P_{22} is a zero matrix, hence if (4.13) may be approximated by a chi-square distribution.

5. Modified and lag order augmented 2SLS estimators

We note that just like the least squares estimator for the VAR model, the application of 2SLS does not provide asymptotically normal or mixed normal estimator because of the long-run endogeneities between lagged I(1) instruments and the (current) shocks of the system. But if we can condition on the innovations driving the common trends it will allow us to establish the independence between Brownian motion of the errors of the conditional system involving the cointegrating relations and the innovations driving the common trends. The idea of the modified 2SLS estimator is to apply the 2SLS method to the equation conditional on the innovations driving the common trends. Unfortunately, the direction of nonstationarity is generally unknown. Neither does the identification condition given by Theorem 2.1 requires such knowledge. In the event that such knowledge is unavailable, Hsiao and Wang [27] propose to generalize Phillips [45] fully modified VAR estimator to the 2SLS estimator.

Rewrite (4.1) as

$$
\begin{aligned}
\underset{\sim}{w}_g &= Z_g\tilde{M}_g\tilde{M}_g^{-1}\underset{\sim}{\delta}_g + \underset{\sim}{\epsilon}_g \\
&= (Z_{g1}^{**} \quad Z_{g2}^{**}) \begin{pmatrix} \underset{\sim}{\delta}_{g1}^{**} \\ \underset{\sim}{\delta}_{g2}^{**} \end{pmatrix} + \underset{\sim}{\epsilon}_g \\
&= Z_g^{**}\underset{\sim}{\delta}_g^{**} + \underset{\sim}{\epsilon}_g
\end{aligned} \tag{5.1}
$$

where $Z_g^{**} = Z_g\tilde{M}_g = (Z_{g1}^{**}, Z_{g2}^{**}), Z_{g1}^{**} = (\bigtriangledown W_g, \bigtriangledown\tilde{W}_{g,-1}, \ldots, \bigtriangledown\tilde{W}_{g,-p+1}), Z_{g2}^{**} = \tilde{W}_{g,-p}, \underset{\sim}{\delta}_g^{**} = \tilde{M}_g^{-1}\underset{\sim}{\delta}_g, \bigtriangledown\tilde{W}_{g,-j}$ denoting the $T \times g_\Delta$ stacked first difference of the included variable $\bigtriangledown\tilde{\underset{\sim}{w}}_{g,t-j}$ and $\bigtriangledown W_g$ denoting the $T \times (g_\Delta - 1)$ first difference of

the included variables $\bigtriangledown\tilde{\underset{\sim}{w}}_{gt}$ excluding $\bigtriangledown \underset{\sim}{w}_{gt}$. The decomposition $(Z_{g1}^{**}, Z_{g2}^{**})$ and $\underset{\sim}{\delta}_g^{**} = (\underset{\sim}{\delta}_{g1}^{**\prime}, \underset{\sim}{\delta}_{g2}^{**\prime})'$ are identical to (Z_{g1}^{*}, Z_{g2}^{*}) if there is no cointegrating relations among $\tilde{\underset{\sim}{w}}_{gt}$, $\underset{\sim}{d}_g = \underset{\sim}{0}$. Unlike $(Z_{g1}^{*}, Z_{g2}^{*}), (Z_{g1}^{**}, Z_{g2}^{**})$ are well defined and observable. When $Z_{g1}^{*} \neq Z_{g1}^{**}$, there exists a nonsingular transformation matrix D_g such that $(Z_{g1}^{**}, Z_{g2}^{**})D_g = (Z_{g1}^{*}, Z_{g2}^{*})$. Then

$$\underset{\sim}{\delta}_g^{*} = D_g^{-1}\underset{\sim}{\delta}_g^{**}. \tag{5.2}$$

Let

$$C_g = (W_{-p}' \bigtriangledown W_{-p} - T\Delta_{\bigtriangledown w \bigtriangledown w})\Omega^{-}_{\bigtriangledown w \bigtriangledown w}\Omega_{\bigtriangledown w \epsilon_g}, \tag{5.3}$$

where Ω_{uv} and Δ_{uv} denote the long-run covariance and the one-sided long-run covariance matrix of two sets of I(0) variables, $(\underset{\sim}{u}_t, \underset{\sim}{v}_t)$,

$$\Omega_{uv} = \sum_{j=-\infty}^{\infty} \Gamma_{uv}(j), \tag{5.4}$$

and

$$\Delta_{uv} = \sum_{j=0}^{\infty} \Gamma_{uv}(j), \tag{5.5}$$

where $\Gamma_{uv}(j) = E\underset{\sim}{u}_t\underset{\sim}{v}'_{t-j}$. Let

$$\hat{C}_g = (W_{-p}' \bigtriangledown W_{-p} - T\hat{\Delta}_{\bigtriangledown w \bigtriangledown w})\hat{\Omega}^{-1}_{\bigtriangledown w \bigtriangledown w}\hat{\Omega}_{\bigtriangledown w \epsilon_g}, \tag{5.6}$$

where $\hat{\Omega}_{uv}$ and $\hat{\Delta}_{uv}$ are the kernel estimates of Ω_{uv} and Δ_{uv}, such as (3.7) and (3.8). A modified 2SLS estimator following Phillips [45] fully modified VAR estimator can be defined as

$$\begin{aligned}\hat{\underset{\sim}{\delta}}_{g,m2SLS}^{**} = &\left\{Z_g^{**\prime}X^{**}(X^{**\prime}X^{**})^{-1}X^{**\prime}Z_g^{**}\right\}^{-1} \\ &\times \left\{Z_g^{**\prime}X^{**}(X^{**\prime}X^{**})^{-1}\begin{pmatrix} X_1^{**\prime}\underset{\sim}{w}_g \\ X_2^{**\prime}\underset{\sim}{w}_g - \hat{C}_g \end{pmatrix}\right\},\end{aligned} \tag{5.7}$$

where $X^{**} = X\tilde{M}_x = (X_1^{**}, X_2^{**}), X_1^{**} = (\bigtriangledown W_{-1}, \ldots, \bigtriangledown W_{-p+1})$, and $X_2^{**} = W_{-p}$. Just like $(Z_{g1}^{**}, Z_{g2}^{**}), (X_1^{**}, X_2^{**})$ are well defined and observable.

Theorem 5.2. *Under assumptions A1-A4, 3.1 and 3.2, the modified 2SLS estimator* $\hat{\underset{\sim}{\delta}}_{g,m2SLS}^{*} = D_g^{-1}\hat{\underset{\sim}{\delta}}_{g,m2SLS}^{**}$ *is consistent. Furthermore*

$$\sqrt{T}(\hat{\underset{\sim}{\delta}}_{g1,m2SLS}^{*} - \underset{\sim}{\delta}_{g1}^{*}) \Longrightarrow N(\underset{\sim}{0}, \sigma_g^2(M_{z_{g1}x_1}^{*}M_{x_1x_1}^{*-1}M_{x_1z_{g1}}^{*})^{-1}) \tag{5.8}$$

and is independent of

$$\begin{aligned}T(\hat{\underset{\sim}{\delta}}_{g2,m2SLS}^{*} - \underset{\sim}{\delta}_{g2}^{*}) \Longrightarrow &(M_{z_{g_2}x_2}^{*}M_{x_2x_2}^{*-1}M_{x_2z_{g_2}}^{*})^{-1} \\ &\cdot M_{z_{g_2}x_2}^{*}M_{x_2x_2}^{*-1}\int B_{x_2^*}dB_{\epsilon_g.x_2^*},\end{aligned} \tag{5.9}$$

which is a mixed normal of the form

$$\int_{M_{x_2^*x_2^*}^{*}>0} N(\underset{\sim}{0}, \sigma^2_{g.\bigtriangledown x_2^*}(M_{z_{g2}x_2}^{*}M_{x_2x_2}^{*-1}M_{x_2z_{g2}}^{*})^{-1})dP(M_{x_2x_2}^{*}). \tag{5.10}$$

where $\sigma^2_{g.\bigtriangledown x_2^*} = \sigma_g^2 - \text{Ł}_{\epsilon_g \bigtriangledown x_2^*}\text{Ł}_{\bigtriangledown x_2^* \bigtriangledown x_2^*}\text{Ł}_{\bigtriangledown x_2^* \epsilon_g}$.

The modified 2SLS estimator of $\underset{\sim}{\delta}_g$ can be obtained as

$$\hat{\underset{\sim}{\delta}}_{g,m2SLS} = \tilde{M}_g \hat{\underset{\sim}{\delta}}^{**}_{g,m2SLS} = \tilde{M}_g D_g \hat{\underset{\sim}{\delta}}^{*}_{g,m2SLS}, \tag{5.11}$$

where $\tilde{M}_g$ is a known matrix but in general, not D_g. However, although the modified 2SLS estimator of $\underset{\sim}{\delta}^*_g$ is either asymptotically normal or mixed normal, the Wald type test statistic

$$\frac{1}{\sigma_g^2}(P\hat{\underset{\sim}{\delta}}_{g,m2SLS} - \underset{\sim}{c})'\{P[Z_g'X(X'X)^{-1}X'Z_g]P'\}^{-1}(P\hat{\underset{\sim}{\delta}}_{g,m2SLS} - \underset{\sim}{c}) \tag{5.12}$$

does not always have the asymptotic chi-square distribution under the null hypothesis $P\underset{\sim}{\delta}_g = \underset{\sim}{c}$, where P is a known $k \times g_\Delta$ matrix of rank k. To see this, rewrite (5.12) in terms of $\hat{\underset{\sim}{\delta}}^*_{g,m2SLS}$

$$\begin{aligned}&\frac{1}{\sigma_g^2}(P^*H_g\hat{\underset{\sim}{\delta}}^*_{g,m2SLS} - \underset{\sim}{c})'\left\{P^*H_g[Z_g^{*'}X^*(X^{*'}X^*)^{-1}X^{*'}Z_g^*]H_g'P^{*'}\right\}\\ &\quad\times(P^*H_g\hat{\underset{\sim}{\delta}}^*_{g,m2SLS} - \underset{\sim}{c}),\end{aligned} \tag{5.13}$$

where $P^* = P\tilde{M}_g D_g H_g^{-1}$ and $H_g = \begin{bmatrix} T^{-1/2}I_{l_g} & 0 \\ 0 & T^{-1}I_{b_g}\end{bmatrix}$. The null hypothesis becomes $P^*H_g\underset{\sim}{\delta}^*_g = \underset{\sim}{c}$. Notice that the asymptotic covariance matrix of $H_g\hat{\underset{\sim}{\delta}}^*_{g,m2SLS}$ converges to

$$\begin{pmatrix}\sigma_g^2(M^*_{z_{g1}'x_1}M^{*-1}_{x_1x_1}M^*_{x_1z_{g1}})^{-1} & \underset{\sim}{0}\\ \underset{\sim}{0} & \sigma^2_{g\cdot\bigtriangledown x_2^*}(M^*_{z_{g2}x_2}M^{*-1}_{x_2x_2}M^*_{x_2z_{g2}})^{-1}\end{pmatrix},$$

while $H_g[Z_g^*X^*(X^{*'}X^*)^{-1}X^{*'}Z_g^*]H_g'$ in (5.13) converges to

$$\sigma_g^2\begin{pmatrix}(M^*_{z_{g1}x_1}M^{*-1}_{x_1x_1}M^*_{x_1z_{g1}})^{-1} & \underset{\sim}{0}\\ \underset{\sim}{0} & (M^*_{z_{g2}x_2}M^{*-1}_{x_2x_2}M^*_{x_2z_{g2}})^{-1}\end{pmatrix}. \tag{5.14}$$

Wald statistic (5.12) (or equivalently (5.13)) is asymptotically chi-square distributed with k degrees of freedom if and only if $P\hat{\underset{\sim}{\delta}}_{g,m2SLS}$ (or equivalently $P^*H_g\hat{\underset{\sim}{\delta}}^*_{g,m2SLS}$) in the hypothesis does not involve the T-consistent component $\hat{\underset{\sim}{\delta}}^*_{g2,m2SLS}$. Otherwise, $H_g[Z_g^{*'}X^*(X^{*'}X^*)^{-1}X^{*'}Z_g^{*'}]H_g'$ would overestimate the asymptotic covariance matrix of $H_g\hat{\underset{\sim}{\delta}}^*_{g,m2SLS}$ because $\sigma^2_{g\cdot\bigtriangledown x_2^*} \le \sigma_g^2$ for the submatrix corresponding to $\underset{\sim}{x}_2^*$ and $\underset{\sim}{z}^*_{g2}$. In general, the test statistic (5.12) is a conservative test, with its asymptotic distribution a weighted sum of k independent χ_1^2 variables with weights between 0 and 1.

The construction of the modified 2SLS estimator requires nonparametric estimation of the long-run covariance matrix and the one-sided long-run covariance matrix. It is well known that kernel estimator and hence the finite sample performance of the modified 2SLS estimator could be affected substantially by the choice of the bandwidth parameter. In addition, since we can not approximate the asymptotic covariance matrix of the modified 2SLS estimator properly, Wald test statistics based on the modified 2SLS estimator using the formula of (5.12) may not be chi-square distributed and critical values that are based on chi-square distributions can be used for conservative tests only. However, as noted by Toda and Yamamoto [55], if we augment the order of a p-th order autoregressive process by

the maximum order of integration then the miscentering and skewness of the limiting distribution of the least squares estimator will be concentrated on the coefficient matrices associated with the augmented lagged vectors which are known *a priori* to be zero, therefore can be ignored. Standard inference procedure can still be applied to the coefficients of the first p coefficient matrices. Hsiao and Wang [28] follow this idea by proposing a lag augmented 2SLS.

The p-th order structural VAR (2.1) can be written as a $(p+1)$-th order structural VAR,

$$A_0\underset{\sim}{w}_t + A_1\underset{\sim}{w}_{t-1} + \cdots + A_p\underset{\sim}{w}_{t-p} + A_{p+1}\underset{\sim}{w}_{t-p-1} = \underset{\sim}{\epsilon}_t, \tag{5.15}$$

where $A_{p+1} \equiv \underset{\sim}{0}$. Transforming (5.15) into an error-correction form, we have

$$\sum_{j=0}^{p} A_j^* \bigtriangledown \underset{\sim}{w}_{t-j} + A_{p+1}^*\underset{\sim}{w}_{t-p-1} = \underset{\sim}{\epsilon}_t, \tag{5.16}$$

where $A_j^* = \sum_{\ell=0}^{j} A_\ell, j = 0, 1, \ldots, p$ and $A_{p+1}^* = A_p^*$. It follows that $A = [A_0, \ldots, A_p] = [A_0^*, \ldots, A_p^*]\tilde{M}^{-1}$.

Let the g-th equation of (5.15) be written as

$$\underset{\sim}{w}_g = Z_g^A\underset{\sim}{\delta}_g^A + \underset{\sim}{\epsilon}_g, \tag{5.17}$$

where $Z_g^A = (Z_g, \tilde{\underset{\sim}{w}}_{g,-(p+1)}), \underset{\sim}{\delta}_g^A = (\underset{\sim}{\delta}'_g, -\underset{\sim}{a}'_{g,p+1})'$ with $\tilde{\underset{\sim}{w}}_{g,-(p+1)}$ denoting the $T \times g_\Delta$ vector of included $\tilde{\underset{\sim}{w}}_{gt}$ lagged by $(p+1)$ periods and $\underset{\sim}{a}_{g,p+1}$ is the g-th row of A_{p+1} excluding those elements subject to exclusion restrictions. Just like (4.1), there exists a nonsingular transformation matrix M_g^A that transforms Z_g^A into $Z_g^{*A} = Z_g^A M_g^A = (Z_{g1}^{*A}, Z_{g2}^{*A})$, and $\underset{\sim}{\delta}_g^{*A} = (M_g^A)^{-1}\underset{\sim}{\delta}_g^A = (\underset{\sim}{\delta}_{g1}^{*A'}, \underset{\sim}{\delta}_{g2}^{*A'})'$ where $Z_{g1}^{*A} = (\bigtriangledown Z_g, \tilde{W}_{g,-(p+1)}\underset{\sim}{\pi}_g)$ is stationary and $Z_{g2}^{*A} = \tilde{W}_{g2,-(p+1)}$ consists of T observed b_g linearly independent $I(1)$ variables, $\tilde{\underset{\sim}{w}}_{g2,t-(p+1)}$. Rewrite (5.17) in terms of the transformed variables,

$$\underset{\sim}{w}_g = Z_g^A M_g^A (M_g^A)^{-1}\underset{\sim}{\delta}_g^A + \underset{\sim}{\epsilon}_g = (Z_{g1}^{*A} \quad Z_{g2}^{*A}) \begin{pmatrix} \underset{\sim}{\delta}_{g1}^{*A} \\ \\ \underset{\sim}{\delta}_{g2}^{*A} \end{pmatrix} + \underset{\sim}{\epsilon}_g \tag{5.18}$$

Let $X^A = (X, W_{-(p+1)})$. The 2SLS estimator of (5.17) is defined as

$$\hat{\underset{\sim}{\delta}}_{g,2SLS}^A = [Z_g^{A'}X^A(X^{A'}X^A)^{-1}X^{A'}Z_g^A]^{-1}[Z_g^{A\prime}X^A(X^{A'}X^A)^{-1}X^{A'}\underset{\sim}{w}_g]. \tag{5.19}$$

The LA2SLS of (4.1) is defined as

$$\hat{\underset{\sim}{\delta}}_{g,LA2SLS} = Q_g^A\hat{\underset{\sim}{\delta}}_{g,2SLS}^A, \tag{5.20}$$

where $Q_g^A = (I_{(p+1)g_\Delta - 1}, \underset{\sim}{0}_{g_\Delta})$, where $\underset{\sim}{0}_{g_\Delta}$ denotes a $[(p+1)g_\Delta - 1] \times g_\Delta$ matrix of zeros. Since $\hat{\underset{\sim}{\delta}}_{g,2SLS}^A = M_g^A\hat{\underset{\sim}{\delta}}_{g,2SLS}^{*A}$, we have

$$\begin{aligned} \hat{\underset{\sim}{\delta}}_{g,LA2SLS} &= Q_g^A M_g^A\hat{\underset{\sim}{\delta}}_{g,2SLS}^{*A} \\ &= (\tilde{M}_g, \underset{\sim}{0}_{g_\Delta})\hat{\underset{\sim}{\delta}}_{g,2SLS}^{*A} \\ &= (\tilde{M}_g, \underset{\sim}{0}_g)\hat{\underset{\sim}{\delta}}_{g1,2SLS}^{*A}, \end{aligned} \tag{5.21}$$

where $\tilde{M}_g$ is a $[(p+1)g_\Delta - 1] \times [(p+1)g_\Delta - 1]$ matrix of the form,[8]

$$\tilde{M}_g = \begin{pmatrix} I_{g_\Delta - 1} & \underset{\sim}{0} & \cdots & \cdots & \underset{\sim}{0} \\ \begin{pmatrix} -I_{g_\Delta - 1} \\ \underset{\sim}{0}' \end{pmatrix} & I_{g_\Delta} & \cdots & \cdots & \cdots \\ \cdots & -I_{g_\Delta} & I_{g_\Delta} & \cdots & \cdots \\ \cdots & \cdots & \cdots & I_{g_\Delta} & \underset{\sim}{0} \\ \cdots & \cdots & \cdots & -I_{g_\Delta} & I_{g_\Delta} \end{pmatrix}, \tag{5.22}$$

with w_{gt} being put as the last element of $\tilde{\underset{\sim}{w}}_{gt}$, I_{g_Δ} denoting the identity matrix of the dimension of included variables in the g-th equation, and $\underset{\sim}{0}_g$ is a $[(p+1)g_\Delta - 1] \times r_g$ matrix with r_g denoting the number of cointegrating relations among $\tilde{\underset{\sim}{w}}_{gt}$ such that $\underset{\sim}{w}'_{gt}\underset{\sim}{\pi}_g$ is $I(0)$. Then $\underset{\sim}{\delta}_g = (\tilde{M}_g, \underset{\sim}{0}_g)\underset{\sim}{\delta}^{*A}_{g1}$.

Since

$$\sqrt{T}(\hat{\underset{\sim}{\delta}}^{*A}_{g1,2SLS} - \underset{\sim}{\delta}^{*A}_{g1}) \longrightarrow N[\underset{\sim}{0}, \sigma_g^2(M^{A*}_{z_{g1}x_1} M^{A*-1}_{x_1x_1} M^{A*}_{x_1z_{g1}})^{-1}], \tag{5.23}$$

where $M^{A*}_{z_{g1}x_1} = \operatorname{plim} \frac{1}{T} Z^{*A'}_{g1} X^{*A}_1, M^{A*}_{x_1x_1} = \operatorname{plim} \frac{1}{T} X^{*A'}_1 X^{*A}_1$, with $X^{*A}_1 = (\bigtriangledown X, W_{-(p+1)}\underset{\sim}{d})$ being the $T \times (mp + r)$ linearly independent $I(0)$ variables. It follows that

Theorem 5.3. *The LA2SLS of $\underset{\sim}{\delta}_g$ is consistent and*

$$\begin{aligned} &\sqrt{T}(\hat{\underset{\sim}{\delta}}_{g,LA2SLS} - \underset{\sim}{\delta}_g) \\ &\quad \Longrightarrow N\left\{\underset{\sim}{0}, \sigma_g^2(\tilde{M}_g \quad \underset{\sim}{0}_g)[M^{A*}_{z_{g1}x_1} M^{A*-1}_{x_1x_1} M^{A*}_{x_1z_{g1}}]^{-1} \begin{pmatrix} \tilde{M}'_g \\ \underset{\sim}{0}'_g \end{pmatrix}\right\}. \end{aligned} \tag{5.14}$$

The LA2SLS estimators of the coefficients of the original structural VAR model (2.1) converge to the true value at the speed of $T^{1/2}$ and are asymptotically normally distributed with nonsingular covariance matrix. Therefore, Wald type test statistics based on LA2SLS estimates are asymptotically chi-square distributed. Compared to the conventional 2SLS or modified 2SLS, the LA2SLS estimator loses the T-convergence component and ignores the prior restrictions that the coefficients on $\tilde{\underset{\sim}{w}}_{g,t-(p+1)}$ are zero, hence may lose some efficiency. However, since distribution of $\hat{\underset{\sim}{\delta}}_g$ is a linear combination of $\hat{\underset{\sim}{\delta}}^*_{g1}$ and $\hat{\underset{\sim}{\delta}}^*_{g2}$ and the limiting distribution of $\hat{\underset{\sim}{\delta}}_{g,LA2SLS}$ is given by the components of the slower rate of convergence, the loss of efficiency in estimating $\hat{\underset{\sim}{\delta}}_g$ by LA2SLS may not be that significant, as reported in a Monte Carlo Study by Hsiao and Wang [28].

6. Conclusions

As demonstrated by Nelson and Plosser [41] that many economic time series are nonstationary. The advancement of nonstationary time series analysis provides a rich reportoire of analytic tools for economists to analyze how do variables respond dynamically to shocks through the decomposition a dynamic system into long-run and short-run relations and allow economists to extract common stochastic trends present in the system that provide information on the important sources of economic fluctuation (e.g. Banerjee, Dolado, Galbraith and Hendry [5], King, Plosser, Stock and Watson [33]). However nonstationarity does not invalid the main concerns of

[8] For ease of notation, we assume all the included variables appear with the same lag order.

Cowles Commission structural approach — identification and simultaneity bias. As shown by Hsiao [26], whether the data is stationary or nonstationary, the same rank condition holds for the identification of an equation in a system. Ignoring the correlations between the regressors and the errors of the equation that arise from the joint dependency of economic variables can lead to severe bias in the least squares estimator even though the regressors are I(1) (Hsiao [21], also see the Monte Carlo study by Hsiao and Wang [28]). Instrumental variable methods have to be applied to obtain consistency.

However, nonstationarity does raise the issue of statistical inference. Standard instrumental variable method can lead to estimators that have non-normal asymptotic distributions and are asymptotically biased and skewed. If there exists prior knowledge to dichotomize the set of variables into joint dependent and exogenous variables and the nonstationarity in the dependent variables is driven by the nonstationarity in the exogenous variables through cointegration relations, standard 2SLS developed for the stationary data can also be used for the analysis of nonstationary data (Hsiao [21, 22]). Wald test statistics for the null are asymptotically chi-square distributed. There is no inference issue. On the other hand, if all the variables are treated as joint dependent as in the time series context, although 2SLS is consistent, the limiting distribution is subject to miscentering and skewness associated with the unit root distribution. Modified or lag order augmented 2SLS will have to be used to ensure valid inference. The modified 2SLS is asymptotically more efficient. However, it also suffers more size distortion in finite sample. On the other hand, the lag order augmented 2SLS does not suffer much efficiency loss, at least in a small scale SVAR model (e.g. Hsiao and Wang [28]), and chi-square distribution is a good approximation for the test statistic.

All above discussions were based on the assumption that no knowledge of cointegration or direction of nonstationarity is known *a priori*. If such information is available, (e.g. King, Plosser, Stock and Watson [33]) estimators incorporating the knowledge of the rank of cointegration presumably will not only lead to efficient estimators of structural form parameters, but also avoid the inference issues arising from the matrix unit roots distrubutions in the system. Unfortunately, structural form estimation methods incorporating reduced rank restrictions appear to be fairly complicated.

The focus of this review is to take a SVAR model as a maintained hypothesis, search for better estimators and understand their properties. We have not looked at the issues of modeling strategy. There is a vast literature on the interactions between structural and non-structural time series analysis to uncover the data-generation process, including testing, estimation, model-combining and prediction (e.g. Hendry and Ericsson [16], Hendry and Krolzig [17], King, Plosser, Stock and Watson [33], Zellner and Palm [61]).

Acknowledgments

I would like to thank Peter Robinson, Arnold Zellner and a referee for helpful comments.

References

[1] Ahn, S. K. and Reinsel, G. C. (1990). Estimation for partially nonstationary autoregressive models. *Journal of the American Statistical Association* **85** 813–823.

[2] Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.
[3] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* **22** 327–351. (Correction, *Annals of Statistics*, 8 (1980), 1400).
[4] Anderson, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Annals of Mathematical Statistics* **30** 676–687.
[5] Banerjee, A., Dolado, J., Galbraith, J. W. and Hendry, D. F. (1993). *Cointegration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford University Press, Oxford.
[6] Browne, M. W. and Arminger, G. (1995). Specification and estimation of mean and covariance structural models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. by G. Arminger, C. C. Clogg and M. Z. Sobel. Plenum, New York, pp. 185–250.
[7] Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *The Annals of Statistics* **16** 367–401.
[8] Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74** 427–431.
[9] Dolado, J. J. and Lutkepohl, H. (1996). Making wald tests work for cointegrated VAR Systems. *Econometric Reviews* **15** 369–386.
[10] Elliott, G. and Stock, J. H. (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory*.
[11] Engle, R. F. and Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica* **55** 251–276.
[12] Gonzalo, J. and Pitarakis, J. Y. (1999). Dimensionality effect in cointegrated systems. In *Granger Festchrift*, ed. by R. Engle and H. White. Oxford University Press, Oxford.
[13] Granger, C. W. J. and Lin, J. L. (1994), Causality in the long-run. *Econometric Theory* **11** 530–536.
[14] Granger, C. W. J., Lin, J. L. and Weiss, A. A. (1983). Time series analysis of error-correction models. In *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. by S. Karlin, T. Amemiya and L. Goodman. Academic Press, New York, pp. 255–278.
[15] Hannan, E. J. (1970). *Multiple Time Series*. Wiley, New York.
[16] Hendry, D. F. and Ericsson, N. R. (1991). An econometric analysis of U.K. money demand in monetary trends in the United States and the United Kingdom, by Milton Friedman and Anna J. Schwartz. *American Economic Review* **81** 8–38.
[17] Hendry, D. F., Ericsson N. R. and Krolzig, H. M. (1999). Improving on data mining. Reconsidered, by K. D. Hoover and S. J. Perez, *Econometrics Journal* **2** 41–58.
[18] Ho, M. and Sorensen, B. (1996). Finding cointegration rank in high dimensional systems using the Johansen Test: An illustration using data based on Monte Carlo simulations. *Review of Economics and Statistics* **78** 726–32.
[19] Hood, W. C. and Koopmans, T. C., eds. (1950). *Studies in Econometric Method*, John Wiley and Sons, New York.

[20] Hosoya, Y. (1996). Causal analysis and statistical inference on possibly non-stationary time series. In D. Kreps and K. F. Wallis eds., *Advances in Economics and Econometrics, 7th World Congress of Econometric Society*, Vol. 3. Cambridge University Press, Cambridge.

[21] Hsiao, C. (1979). Autoregressive modelling of Canadian money and income data. *Journal of the American Statistical Association* **74** 553–560.

[22] Hsiao, C. (1979). Causality Tests in Econometrics. *Journal of Economic Dynamics and Control* **1** 321–46.

[23] Hsiao, C. (1983). Identification. In *Handbook of Econometrics*, vol. 1. ed. by Z. Griliches and M. Intriligator. North-Holland, Amsterdam, pp. 223–283.

[24] Hsiao, C. (1997). Cointegration and dynamic simultaneous equations models. *Econometrica* **65** 647–670.

[25] Hsiao, C. (1997). Statistical properties of the two stage least squares estimator under cointegration. *Review of Economic Studies* **64** 385–398.

[26] Hsiao, C. (2001). Identification and dichotomization of long- and short-run relations of cointegrated vector autoregressive models. *Econometric Theory* **17** 889–912.

[27] Hsiao, C. and Wang, S. (2006). Modified two stage least squares estimator for the estimation of a structural vector autoregressive integrated process. *Journal of Econometrics* **135** 427–463.

[28] Hsiao, C. and Wang, S. (2006). A lag augmented two and three stage least squares estimator for structural vector autoregressive integrated processes. *Econometrics Journal*, forthcoming.

[29] Hsiao, C., Shen, Y. and Fujiki, H. (2005). Aggregate vs. disaggregate data analysis — A paradox in the estimation of money demand function of Japan under the low interest rate policy. *Journal of Applied Econometrics* 579–601.

[30] Intriligator, M. D., Bodkin, R. G. and Hsiao, C. (1996). *Econometric Models, Techniques, and Applications*, 2nd edn. Prentice-Hall, Upper Saddle River, NJ.

[31] Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12** 231–254.

[32] Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59** 1551–1580.

[33] King, R., Plosser, C. I., Stock, J. H. and Watson, M. W. (1991). Stochastic trends and economic fluctuation. *American Economic Review* **81** 819–840.

[34] Klein, L. R. (1950). *Economic Fluctuations in the United States, 1921-1941.* Cowles Commission Monograph **11**. Wiley and Sons, Inc., New York.

[35] Klein, L. R. and Evans, M. K. (1969). *Econometric Gaming.* MacMillan, New York.

[36] Klein, L. R. and Goldberger, A. S. (1955). *An Econometric Model of the United States*, North Holland, Amsterdam.

[37] Koopmans, T. C., Rubin, H. and Leipnik, R. B. (1950). Measuring the equation systems of dynamic economics. In *Statistical Inference in Dynamic Economic Models*, ed. by T.C. Koopman. Wiley, New York, pp. 53–237.

[38] Malinvaud, E. (1980). *Statistical Methods of Econometrics*, 3rd edn. North-Holland, Amsterdam.

[39] Muthén, B. (1984). A general structural equation model with Dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49** 115–132.

[40] Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika* **29** 81–177.
[41] Nelson, C. R. and Plosser, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* **10** 139–162.
[42] Park, J. Y. (1992). Canonical cointegrating regressions. *Econometrica* **60** 119–143.
[43] Pesaran, M. H., Shin, Y. and Smith, R. (2000). Structural analysis of vector error-correction models with exogenous $I(1)$ variables. *Journal of Econometrics* **97** 293–343.
[44] Phillips, P. C. B. (1991). Optimal inference in cointegrating systems. *Econometrica* **59** 283–306.
[45] Phillips, P. C. B. (1995). Fully modified least squares and vector autoregression. *Econometrica* **63** 1023–1078.
[46] Phillips, P. C. B. and Durlauf, S. N. (1986). Multiple time series regression with integrated processes. *Review of Economic Studies* **53** 473–495.
[47] Phillips, P. C. B. and Hansen, B. E. (1990). Statistical inference in instrumental variables regression with $I(1)$ processes. *Review of Economic Studies* **57** 99–125.
[48] Priestley, M. B. (1982). *Spectral Analysis and Time Series, Vols. I and II.* Academic Press, New York.
[49] Robinson, P. M. and Hualda, J. (2003). Cointegration in fractional systems with unknown integration orders. *Econometrica* 1727–1766.
[50] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.
[51] Sims, C. A., Stock, J. H. and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica* **58** 113–144.
[52] Stock, J. H. (1987). Asymptotic properties of least squares estimators of cointegrating vectors. *Econometrica* **55** 1035–1056.
[53] Stock, J. H. (2001). Macro-econometrics. *Journal of Econometrics* **100** 29–32.
[54] Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *Journal of the American Statistical Association* **76** 802–816.
[55] Toda, H. Y. and Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics* **66** 225–250.
[56] Toda, H. Y. and Phillips, P. C. B. (1993). Vector autorregression and caussality. *Econometrica* **61** 1367–1393.
[57] Tsay, R. S. and Tiao, G. C. (1990). Asymptotic properties of multivariate nonstationary processes with applications to autoregressions. *Annals of Statistics* **18** 220–250.
[58] Watson, M. W. (1994). Vector autoregressions and cointegration. In *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. L. McFadden. North-Holland, Amsterdam, pp. 2844–2915.
[59] Yuan, K. H. and Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association* **92** 767–774.
[60] Zellner, A. and Palm, F. (1974). Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* **2** 17–54.
[61] Zellner, A. and Palm, F. (2004). *The Structural Econometric Time Series Approach.* Cambridge University Press, Cambridge.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 193–209

DOI: 10.1214/074921706000001049

Combining domain knowledge and statistical models in time series analysis

Tze Leung Lai[1] **and Samuel Po-Shing Wong**[2]

Stanford University and The Chinese Universty of Hong Kong

Abstract: This paper describes a new approach to time series modeling that combines subject-matter knowledge of the system dynamics with statistical techniques in time series analysis and regression. Applications to American option pricing and the Canadian lynx data are given to illustrate this approach.

1. Introduction

In their Fisher Lectures at the Joint Statistical Meetings, Cox [11] and Lehmann [31] mentioned two major types of stochastic models in statistical analysis, namely, *empirical* and *substantive* (or mechanistic). Whereas substantive models are explanatory and related to subject-matter theory on the mechanisms generating the observed data, empirical models are interpolatory and aim to represent the observed data as a realization of a statistical model chosen largely for its flexibility, tractability and interpretability but not on the basis of subject-matter knowledge. Cox [11] also mentioned a third type of stochastic models, called *indirect* models, that are used to evaluate statistical procedures or to suggest methods for analyzing complex data (such as hidden Markov models in image analysis). He noted, however, that the distinctions between the different types of models are important mostly when formulating and checking them but that these types are not rigidly defined, since "quite often parts of the model, e.g., those representing systematic variation, are based on substantive considerations with other parts more empirical." In this paper, we elaborate further the complementary roles of empirical and substantive models in time series analysis and describe a basis function approach to combining subject-matter (domain) knowledge with statistical modeling techniques.

This basis function approach was first developed in [29] for the valuation of American options. In Sections 2 and 3 we review the statistical and subject-matter models for option pricing in the literature as examples of empirical and substantive models in time series analysis. Section 4 describes a combined substantive-empirical approach via basis functions, in which the substantive component is associated with basis functions of a certain form, and the empirical component uses flexible and computationally convenient basis functions such as regression splines. The work of Lai and Wong [29] on option pricing and recent related work in financial time series are reviewed to illustrate this approach. Section 5 applies this approach to a widely studied data set in the nonlinear time series literature, namely, the Canadian

[1]Department of Statistics, Stanford Univeristy, Stanford, CA 94305, U.S.A., e-mail: lait@stat.stanford.edu

[2]Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, e-mail: samwong@sta.cuhk.edu.hk

AMS 2000 subject classifications: primary 62M10, 62M20; secondary 62P05, 62P10.

Keywords and phrases: time series analysis, domain knowledge, empirical models, mechanistic models, combined substantive-empirical approach, basis function.

lynx data set that records the annual numbers of Canadian lynx trapped in the Mackenzie River district from 1821 to 1934. We use substantive models from the ecology literature together with multivariate adaptive regression splines to come up with a new time series model for these data. Some concluding remarks are given in Section 6.

2. Statistical (empirical) time series models

The development of statistical time series models in the past fifty years has witnessed a remarkable confluence of basic ideas from various areas in statistics and probability, coupled with the powerful influence from diverse fields of applications ranging from economics and finance to signal processing and control systems. The first phase of this development was concerned with stationary time series, leading to MA (moving average), AR (autoregressive) and ARMA representations in the time domain and transfer function representations in the frequency domain. This was followed by extensions to nonstationary time series, either by fitting (not necessarily stationary) ARMA models or by the Box-Jenkins approach involving the ARIMA (autoregressive integrated moving average) models and their seasonal SARIMA counterparts. More general fractional differencing then led to the ARFIMA models. The next phase of the development was concerned with nonlinear time series models, beginning with bilinear models that add cross-product terms $y_{t-i}\epsilon_{t-j}$ to the usual ARMA model $y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \epsilon_t + c_1\epsilon_{t-1} + \cdots + c_q\epsilon_{t-q}$, and threshold autoregressive and regime switching models that introduce nonlinearities into the usual autoregressive models via state-dependent changes or Markov jumps in the autoregressive parameters. The monograph by Tong [44] summarized these and other nonlinear time series models in the previous literature. The appropriateness of the parametric forms assumed in these nonlinear time series models, however, may be difficult to justify in real applications, as pointed out by Chen and Tsay [9].

Whereas the AR model $y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \epsilon_t$ is related to linear regression since $\boldsymbol{\beta}^T \mathbf{x}_t$ is the regression function $E(y_t|\mathbf{x}_t)$ of y_t given $\mathbf{x}_t := (y_{t-1}, \ldots, y_{t-p})^T$, and likewise its nonlinear parametric extensions $y_t = f(\mathbf{x}_t, \boldsymbol{\beta}) + \epsilon_t$ are related to nonlinear regression, Chen and Tsay [9, 10] proposed to use nonparametric regression for $E(y_t|\mathbf{x}_t)$ instead. They started with functional-coefficient autoregressive (FAR) models of the form $y_t = f_1(\mathbf{x}_t^*)y_{t-1} + \cdots + f_p(\mathbf{x}_t^*)y_{t-p} + \epsilon_t$, where $f_1, \ldots, f_p$ are unspecified functions to be estimated by local linear regression and $\mathbf{x}_t^* = (y_{t-i_1}, \ldots, y_{t-i_d})^T$ with $i_1 < \cdots < i_d$ chosen from $\{1, \ldots, p\}$. Because of sparse data in high dimensions, local linear regression typically require d to be 1 or 2. To deal with nonparametric regression in higher dimensions, they considered additive autoregressive models of the form $y_t = f_1(y_{t-i_1}) + \cdots + f_d(y_{t-i_d}) + \epsilon_t$, in which the f_i can be estimated nonparametrically via the generalized additive model (GAM) of Hastie and Tibshirani [19] . Making use of Friedman's [15] multivariate adaptive splines (MARS), Lewis and Stevens [34] and Lewis and Ray [32, 33] developed spline models for empirical modeling of time series data. Weigend, Rummelhart and Huberman [48] and Weigend and Gershenfeld [47] proposed to use neural networks (NN) to model $E(y_t|\mathbf{x}_t)$, while Lai and Wong [28] considered a variant called stochastic neural networks, for which they could use the EM algorithm to develop efficient estimation procedures that have much lower computational complexity than those for conventional neural networks.

The preceding time series models are autonomous, relating the dynamics of y_t to

the past states. In econometrics and engineering, the outputs y_t are related not only to the past outputs but also to the past inputs $u_{t-d}, \ldots, u_{t-k}$. Therefore the AR model has been extended to the ARX model (where X stands for exogenous inputs) $y_t = \boldsymbol{\beta}^T \mathbf{x}_t + \epsilon_t$ with $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-p}, u_{t-d}, \ldots, u_{t-k})^T$. Instead of assuming a linear or nonlinear parametric regression model, one can use nonparametric regression to estimate $E(y_t|\mathbf{x}_t)$, as in the following financial application.

Example 1. As noted by Ross [40], option pricing theory is "the most successful theory not only in finance, but in all of economics." A call (put) option gives the holder the right to buy (sell) the underlying asset (e.g. stock) by a certain date T (known as the "expiration date" or "maturity") at a certain price (known as the "strike price" and denoted by K). European options can be exercised only on the expiration date, whereas American options can be exercised at any time up to the expiration date. The celebrated Black-Scholes theory, which will be reviewed in Section 3, yields the following pricing formulas for the prices c_t and p_t of European call and put options at time $t \in [0, T)$:

$$c_t = S_t e^{-d(T-t)} \Phi(d_1(S_t, K, T-t)) - K e^{-r(T-t)} \Phi(d_2(S_t, K, T-t)), \tag{2.1}$$

$$p_t = K e^{-r(T-t)} \Phi(-d_2(S_t, K, T-t)) - S_t e^{-d(T-t)} \Phi(-d_1(S_t, K, T-t)), \tag{2.2}$$

where Φ is the cumulative distribution function of the standard normal random variable, S_t is the price of the underlying asset at time t, d is the dividend rate of the underlying asset, $d_1(x, y, v) = \{\log(x/y) + (r - d + \sigma^2/2)v\}/\sigma\sqrt{v}$ and $d_2(x, y, v) = d_1(x, y, v) - \sigma\sqrt{v}$. Hutchinson, Lo and Poggio [22] pointed out that the success of the formulas (2.1) and (2.2) depends heavily on the specification of the dynamics of S_t. Instead of using any particular model of S_t, they proposed a data-driven way for pricing and hedging with a minimal assumption: independent increments of the underlying asset price. Noting that y_t ($= c_t$ or p_t) is function of S_t/K and $T - t$ with r and σ being constant, they assume $y_t = Kf(S_t/K, T-t)$ and approximate f by taking $\mathbf{x}_t = (S_t/K, T-t)^T$ in the following models:

(i) radial basis function (RBF) networks $f(\mathbf{x}) = \beta_0 + \boldsymbol{\alpha}^T \mathbf{x} + \sum_{i=1}^I \beta_i h_i(\|A(\mathbf{x} - \gamma_i)\|)$, where A is a positive definite matrix and h_i is of the RBF type e^{-u^2/σ_i^2} or $(u^2 + \sigma_i^2)^{1/2}$;

(ii) neural networks $f(\mathbf{x}) = \psi(\beta_0 + \sum_{i=1}^I \beta_i h(\boldsymbol{\gamma}_i + \boldsymbol{\alpha}_i^T \mathbf{x}))$, where $h(u) = 1/(1+e^{-u})$ is the logistic function and ψ is either the identity function or the logistic function;

(iii) projection pursuit regression (PPR) networks $f(\mathbf{x}) = \beta_0 + \sum_{i=1}^I \beta_i h_i(\boldsymbol{\alpha}_i^T \mathbf{x})$, where h_i is an unspecified function that is estimated from the data by PPR.

The $\boldsymbol{\alpha}_i$, β_i and $\boldsymbol{\gamma}_i$ above are unknown parameters of the network that are to be estimated from the data. As pointed out in [22], all three classes of networks have some form of "universal approximation property" which means their approximation bounds do not depend on the dimensionality of the predictor variable x; see [2]. It should be noted that the above transformation of S_t to S_t/K can be motivated not only from the assumption on S_t but also from the special feature of options data. Although the strike price K could be any positive number theoretically, the options exchange only sets strike prices at a multiple of a fundamental unit. For example, Chicago Board Options Exchange (CBOE) sets strike prices at multiples of \$5 for stock prices in the \$25 to \$200 range. Also, only those options with strike prices closet to the current stock price are traded and thus their prices are observed. Since S_t is non-stationary in general, the observed K is also non-stationary. Such features

create sparsity of data in the space of $(S_t, K, T-t)$. Training the options pricing formula in the form of $f(S_t, K, T-t)$ can only interpolate the data and can hardly produce any good prediction because (S_t, K) in the future can be very different from the data used in estimating f. The proposed transformation makes use of the fact that all observed and future S_t/K are close to 1. Therefore, the proposed transformation captures the stationary structure of the data and enable the nonparametric models to predict well. Another point that Hutchinson, Lo and Poggio [22] highlighted is the measure of performance of the estimated pricing formula. According to their simulation study, even a linear $f(S_t/K, T-t)$ can give $R^2 \approx 90\%$ (Table I of Hutchinson, Lo and Poggio [22]). However, such a linear f implies a constant delta hedging scheme which would provide poor hedging results. Since the primary function of options is hedging the risk created by changes in the price of the underlying asset, Hutchinson, Lo and Poggio [22] suggested using, instead of R^2, the hedging error measures $\xi = e^{-rT}E[|V(T)|]$ and $\eta = e^{-rT}[EV^2(T)]^{1/2}$, where $V(T)$ is the value of the hedged portfolio at time T. In a perfect Black-Scholes world, $V(T)$ should be 0 if Black-Scholes formula is used. However, from the simulation study, the Black-Scholes formulas still give $\xi > 0$ and $\eta > 0$ because time is discrete. Hutchinson, Lo and Poggio [22] reported that RBF, NN and PPR all give hedging measures comparable to those of the Black-Scholes in the simulation study. For real data analysis of futures options, RBF, NN and PPR performed better than the Black-Scholes formula in terms of hedging.

For American options, instead of using these learning networks to approximate the option price, Broadie et al. [5] used kernel smoothers to estimate the option pricing formula of an American option. Using a training sample of daily closing prices of American calls on the S&P100 Index that were traded on the Chicago Board Options Exchange from 3 January 1984 to 30 March 1990, they compared the nonparametric estimates of American call option prices at a set of $(S/K, t^*)$ values with corresponding parametric estimates obtained by using the approximations to American option prices due to Broadie and Detemple [4], and found significant differences between the parametric and nonparametric estimates.

3. Substantive (mechanistic) models

In control engineering, the dynamics of linear input-output systems are often given by ordinary differential equations, whose discrete-time approximations in the presence of noise have led to the ARX models (for white noise), and ARMAX models (for colored noise) in the preceding section. The problem of choosing the inputs sequentially so that the outputs are as close as possible to some target values when the model parameters are unknown and have to be estimated on-line has a large literature under the rubric of *stochastic adaptive control*; see Goodwin, Ramadge and Caines [16], Lai and Wei [27], Lai and Ying [30] and Guo and Chen [17]. More general dynamics in the presence of additive noise have led to stochastic differential equations (SDEs), whose discrete-time approximations are related to nonlinear time series models described in the preceding section. One such SDE is geometric Brownian motion (GBM) for the asset price process in the Black-Scholes option pricing theory. In view of Ito's formula, the GBM dynamics for the asset price S_t translate into SDE dynamics for the option price $f(t, S_t)$. Such implied dynamics from the mechanistic model can be combined with subject-matter theory to derive the functional form or differential equation for f and other important corollaries of the theory, as illustrated in the following.

Example 2. In the Black-Scholes model, the asset price S_t is assumed to be GBM defined by the SDE

$$dS_t/S_t = \mu dt + \sigma dw_t, \tag{3.1}$$

where $w_t, t \geq 0$, is Brownian motion. Letting $f(t, S)$ be the price of the option at time t when $S_t = S$, it follows from (3.1) and Ito's formula that

$$\begin{aligned} df(t, S_t) &= \tfrac{\partial f}{\partial t} dt + \tfrac{\partial f}{\partial S} dS_t + \tfrac{1}{2} \tfrac{\partial^2 f}{\partial S^2} \sigma^2 S_t^2 dt \\ &= \Big(\tfrac{\partial f}{\partial t} + \mu S_t \tfrac{\partial f}{\partial S} + \tfrac{1}{2} \sigma^2 S_t^2 \tfrac{\partial^2 f}{\partial S^2} \Big) dt + \sigma S_t \tfrac{\partial f}{\partial S} dw_t. \end{aligned}$$

For simplicity assume that the asset does not pay dividends, i.e., $d = 0$. Consider an option writer's portfolio at time t, consisting of -1 option and y_t units of the asset. The value of the portfolio π_t is $-f(t, S_t) + y_t S_t$ and therefore

$$d\pi_t = -\Big(\frac{\partial f}{\partial t} + \mu S_t \frac{\partial f}{\partial S} + \frac{1}{2} \sigma^2 S_t^2 \frac{\partial^2 f}{\partial S} - \mu y_t S_t \Big) dt + \sigma S_t \Big(y_t - \frac{\partial f}{\partial S} \Big) dw_t.$$

Hence setting $y_t = \partial f / \partial S$ yields a risk-free portfolio. This is the basis of *delta hedging* in the options theory of Black and Scholes [3], who denote $\partial f / \partial S$ by Δ. Besides GBM dynamics for the asset price, the Black-Scholes theory also assumes that there are no transaction costs and no limits on short selling and that trading can take place continuously so that delta hedging is feasible. Since economic theory prescribes absence of arbitrage opportunities in equilibrium, π_t that consists of -1 option and Δ units of the asset should have the same return as $r\pi_t dt = r(-f + S_t \Delta) dt$, yielding the Black-Scholes PDE for f:

$$\frac{\partial f}{\partial t} + rS \frac{\partial f}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 f}{\partial S^2} = rf, \quad 0 \leq t < T, \tag{3.2}$$

with the boundary condition $f(T, S) = g(S)$, where $g(S) = (K - S)_+$ for a put option, and $g(S) = (S - K)_+$ for a call option, where $x_+ = \max(x, 0)$. This PDE has the explicit solution (2.1) or (2.2) with $d = 0$. If the asset pays dividend at rate d, then a modification of the preceding argument yields (3.2) in which $rS(\partial f / \partial S)$ is replaced by $(r - d)S(\partial f / \partial S)$.

Merton [37] extended the Black-Scholes theory for pricing European options to American options that can be exercised at any time prior to the expiration date. Optimal exercise of the option is shown to occur when the asset price exceeds (or falls below) an exercise boundary $\partial \mathcal{C}$ for a call (or put) option. The Black-Scholes PDE still holds in the continuation region $\mathcal{C}$ of (t, S_t) before exercise, and $\partial \mathcal{C}$ is determined by the free boundary condition $\partial f / \partial S = 1$ (or -1) for a call (or put) option. Unlike the explicit formula (2.1) or (2.2) for European options, there is no closed-form solution of the free-boundary PDE and numerical methods such as finite differences are needed to compute American option prices under this theory.

By the Feynman-Kac formula, the PDE (3.2) has a probabilistic representation $f(t, S) = E[e^{-r(T-t)} g(S_T) | S_t = S]$, and the expectation E is with respect to the "equivalent martingale measure" under which $dS_t/S_t = (r - d)dt + \sigma dw_t$. This representation generalizes to American options as the value function of the optimal stopping problem

$$f(t, S) = \sup_{\tau \in \mathcal{T}_{t,T}} E[e^{-r(\tau - t)} g(S_\tau) | S_t = S] \tag{3.3}$$

where $\mathcal{T}_{t,T}$ denotes the set of stopping times τ taking values between t and T. Cox, Ross and Rubinstein [12] proposed to approximate GBM by a binomial tree, with root node S_0 at time 0, so that (3.3) can be approximated by a discrete-time and discrete-state optimal stopping problem that can be solved by backward induction. Denote $f(t,S)$ by $C(t,S)$ for an American call option, and by $P(t,S)$ for an American put option. Jacka [23] and Carr, Jarrow and Myneni [7] derived the decomposition formula

$$\begin{aligned} P(t,S) = p(t,S) + K\rho e^{\rho u} \int_u^0 \Big\{ & e^{-\rho s}\Phi\Big(\frac{\bar z(s)-z}{\sqrt{s-u}}\Big) \\ & - \theta e^{-(\theta\rho s + u/2)+z}\Phi\Big(\frac{\bar z(s)-z}{\sqrt{s-u}} - \sqrt{s-u}\Big)\Big\} ds, \end{aligned} \tag{3.4}$$

and a similar formula relating $C(t,S)$ to $c(t,S)$, where $\bar z(u)$ is the early exercise boundary $\partial\mathcal{C}$ under the transformation

$$\rho = r/\sigma^2,\ \theta = d/r;\ u = \sigma^2(t-T),\ z = \log(S/K) - (\rho - \theta\rho - 1/2)u. \tag{3.5}$$

Ju [24] found that the early exercise premium can be computed in closed form if $\partial\mathcal{C}$ is a piecewise exponential function which corresponds to a piecewise linear $\bar z(u)$. By using such assumption, Ju [24] reported numerical studies showing his method with 3 equally spaced pieces substantially improves previous approximations to option prices in both accuracy and speed. AitSahlia and Lai [1] introduced the transformation (3.5) to reduce GBM to Brownian motion and showed that $\bar z(u)$ is indeed well approximated by a piecewise linear function with a few pieces. The integral obtained by differentiating that in (3.4) with respect to S also has a closed-form expression when $\bar z(\cdot)$ is piecewise linear, and approximating $\bar z(\cdot)$ by a linear spline that uses a few unevenly spaced knots gives a fast and reasonably accurate method for computing $\Delta = \partial P/\partial S$.

The Black-Scholes price involves the parameters r and σ, which need to be estimated. The yield of a short-maturity Treasury bill is usually used for r. Although in the GBM model for asset prices which are observed at fixed intervals of time (e.g. daily), one can estimate σ by the standard deviation of historical (daily) asset returns, which are i.i.d. normal under the GBM model for asset prices, there are issues due to departures from this model (e.g., σ can change over time and asset returns are markedly non-normal) and due to violations of the Black-Scholes assumptions in the financial market (e.g., there are actually transaction costs and limits on short selling). Section 13.4 and Chapter 16 of Hull [21] discuss how the parameter σ in the Black-Scholes option price is treated in current practice. In the next section we describe an alternative approach that addresses the discrepancy between the Black-Scholes-Merton theory and time series data on American options and the underlying stock prices.

4. A combined substantive-empirical approach

In this section we describe an approach to time series modeling that contains both substantiative and empirical components. We first came up with this approach when we studied valuation of American options. Its basic idea is to use empirical modeling to address the gap between the actual prices in the American options market and the option prices given by the Black-Scholes-Merton theory in Example 2, as explained below.

Example 3. For European options, instead of using the basis function of Hutchinson, Lo and Poggio [22], an alternative approach is to express the option price as $c + Ke^{-rt^*} f^*(S/K, t^*)$, where c is the Black-Scholes price (2.1) because the Black-Scholes formula has proved to be quite successful in explaining empirical data. This is tantamount to including $c(t, S)$ as one of the basis functions (with prescribed weight 1) to come up with a more parsimonious approximation to the actual option price.

The usefulness of this idea is even more apparent in the case of American options. Focusing on puts for definiteness, the decomposition formula (3.4) expresses an American put option price as the sum of a European put price p and the early exercise premium which is typically small relative to p. This suggests that p should be included as one of the basis functions (with prescribed weight 1). Lai and Wong [29] propose to use additive regression splines after the change of variables $u = -\sigma^2(T-t)$ and $z = \log(S/K)$. Specifically, for small $T-t$ (say within 5 trading days prior to expiration, i.e. $T-t \le 5/253$ under the assumption of 253 trading days per year), we approximate P by p. For $T-t > 5/253$ (or equivalently, $u < -5\sigma^2/253$), we approximate P by

$$\begin{aligned} P = p + Ke^{\rho u}\{&\alpha + \alpha_1 u + \sum_{j=1}^{J_u} \alpha_{1+j}(u - u^{(j)})_+ \\ &+ \beta_1 z + \beta_2 z^2 + \sum_{j=1}^{J_z} \beta_{2+j}(z - z^{(j)})_+^2 + \gamma_1 w + \gamma_2 w^2 \\ &+ \sum_{j=1}^{J_w} \gamma_{2+j}(w - w^{(j)})_+^2\}, \end{aligned} \tag{4.1}$$

where $\rho = r/\sigma^2$ as in (3.5), α, α_j, β_j and γ_j are regression parameters to be estimated by least squares from the training sample and

$$w = |u|^{-1/2}\{z - (\rho - \theta\rho - 1/2)u\} \quad (\theta = d/r) \tag{4.2}$$

is an "interaction" variable derived from z and u. The motivation behind the centering term $(\rho - \theta\rho - 1/2)u$ comes from (3.5) that transforms GBM into Brownian motion, whereas that behind the normalization $|u|^{-1/2}$ comes from (3.4) and the closely related $d_1(x, y, v)$ in (2.2). The knots $u^{(j)}$ (respectively $z^{(j)}$ or $w^{(j)}$) of the linear (respectively quadratic) spline in (4.1) are the $100j/J_u$ (respectively $100j/J_z$ and $100j/J_w$)-th percentiles of $\{u_1, \ldots, u_n\}$ (respectively $\{z_1, \ldots, z_n\}$ or $\{w_1, \ldots, w_n\}$). The choice of J_u, J_z and J_w is over all possible integers between 1 and 10 to minimize the generalized cross validation (GCV) criterion, which can be expressed in the following form (cf. [19, 46]):

$$\mathrm{GCV}(J_u, J_z, J_w) = \sum_{i=1}^{n}(P_i - \hat{P}_i)^2 \Bigg/ \left\{ n\left(1 - \frac{J_u + J_z + J_w + 6}{n}\right)^2 \right\},$$

where the P_i are the observed American option prices in the past n periods, and the $\widehat{P}_i$ are the corresponding fitted values given by (4.1) in which the regression coefficients are estimated by least squares.

In the preceding we have assumed prescribed constants γ and σ as in the Black-Scholes model; these parameters appear in (4.1) via the change of variables (3.5). In practice σ is unknown and may also vary with time. We can replace it in (4.1)

by the standard deviation $\widehat{\sigma}_t$ of the most recent asset returns say, during the past 60 trading days prior to t as in [22], p. 881. Moreover, the risk-free rate r may also change with time, and can be replaced by the yield $\widehat{r}_t$ of a short-maturity Treasury bill on the close of the month before t. The same remark also applies to the dividend rate.

The simulation study in Lai and Wong [29] shows the advantages of this combined substantive-empirical approach. Not only is P well approximated by $\widehat{P}$, especially over intervals of S/K values that occur frequently in the sample, $\widehat{\Delta} - \Delta$ also reveals a pattern similar to that of $\widehat{P} - P$. Besides $\xi_{\hat{P}} = E\{e^{-r\tau}|V_{\hat{P}}(\tau)|\}$, where τ is the time of exercise and $V_{\hat{P}}(t)$ is the value of the replicating portfolio at time t that rebalances (according to the pricing formula $\hat{P}$) between the risky and riskless assets ([22], p. 868-869), Lai and Wong [29] also consider the measure

$$\kappa_{\hat{P}} = E\left\{\int_0^\tau (S_t/K)^2(\Delta(t) - \hat{\Delta}(t))^2 dt\right\}, \tag{4.3}$$

where $\hat{\Delta} = \partial\hat{P}/\partial S$. In practice, continuous rebalancing is not possible. If rebalancing is done only daily, then $(S/K)^2(\Delta_A - \hat{\Delta})^2$ in (4.3) is replaced by a step function that stays constant on intervals of width $1/253$. Because of the adaptive nature of the methodology, the proposed approach of Lai and Wong [29] is much more robust to the misspecification error than the Black-Scholes formula in terms of both measures. Lai and Lim [26] carried out an empirical study of this approach and made use of its semiparametric pricing formula and (4.3) to come up with a modified Black-Scholes theory and optimal delta hedging in the presence of transaction costs.

5. Application to the 1821-1934 Canadian lynx data

The Canadian Lynx data set consists of the annual record of the numbers of the Canadian lynx trapped in the Mackenzie River district of the North-west Canada for the period 1821-1934 inclusively. Let X_t be $\log_{10}$(number recorded as trapped in year $1820+t$) ($t = 1, \ldots, 114$). Figure 1 shows the time series plot of X_t. According to Tong [44], Moran [39] performed the first time series analysis on these data by fitting an AR(2) model to X_t; moreover, the log transformation is used because it (i) makes the marginal distribution of X_t more symmetric about its mean and (ii) reduces the approximation error in assuming the number of lynx to be proportional to the population. In view of the substantial non-linearity of $E[X_t|X_{t-3}]$ found in the scatterplot of X_t versus X_{t-3}, Tong([44], p.361) critiques Moran's analysis and its enhancements by Campbell and Walker [6], who added a harmonic component to the AR(2) model, and by Tong [43], who used the AIC to select the order $p = 11$ for AR(p) models, as "uncritical acceptance of linearity" in X_t. He uses a self-excited threshold autoregressive model (SETAR) of the form

$$X_t - X_{t-1} = \begin{cases} 0.62 + 0.25X_{t-1} - 0.43X_{t-2} + \varepsilon_t & \text{if } X_{t-2} \le 3.25 \\ -(1.24X_{t-2} - 2.25) + 0.52X_{t-1} + \varepsilon_t & \text{if } X_{t-2} > 3.25 \end{cases} \tag{5.1}$$

to fit these data, similar to Tong and Lim ([45], Section 9). The growth rate $X_t - X_{t-1}$ in the first regime (i.e., $X_{t-2} \le 3.25$) tends to be positive but small, which corresponds to a slow population growth. In the second regime (i.e., $X_{t-2} > 3.25$), $X_t - X_{t-1}$ tends to be negative, corresponding to a decrease in population size.

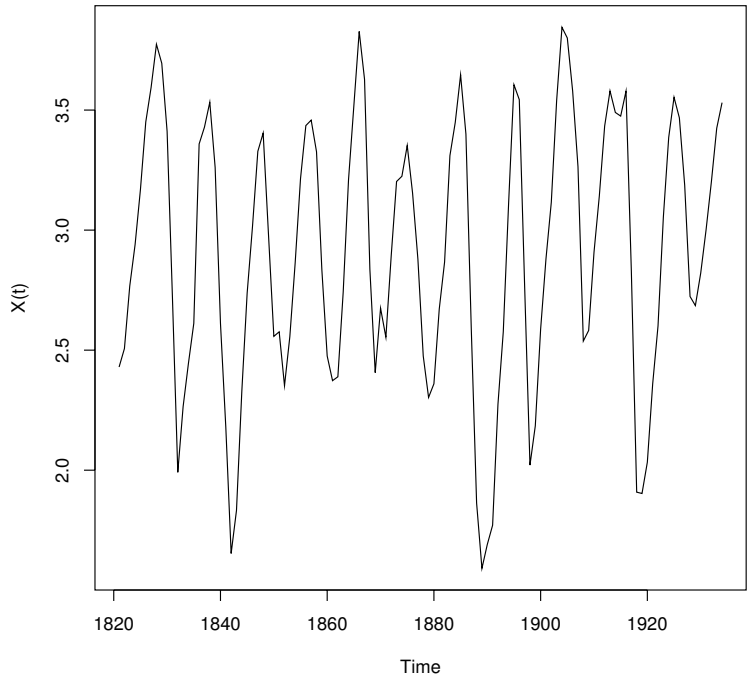


FIG 1. *Time series plot of* $\log_{10}$ *of the Canadian lynx series.*

Tong ([44], p. 377) interprets the fitted model as an "energy balance" between the population expansion and the population contraction, yielding a stable limit cycle with a 9-year period which is in good agreement with the observed asymmetric cycles. Motivated by Van der Pol's equation, Haggan and Ozaki [18] proposed to fit another nonlinear time series model, namely, the exponential autoregressive model

$$X_t - \mu = \sum_{j=1}^{11} (\phi_j + \pi_j e^{-\gamma (X_{t-j}-\mu)^2})(X_{t-j} - \mu) + \varepsilon_t, \tag{5.2}$$

which gives a limit cycle of period 9.45 years. Lim [35] compares the prediction performance of these and other parametric models and concludes that Tong's SETAR model ranks the best among them.

Taking a more nonparametric approach, Fan and Yao [14] use a functional − coefficient autoregressive model to fit the observed X_t series and compare its prediction with that of threshold autoregression. Specifically, they fit the FAR(2,2) model

$$X_t = a_1(X_{t-2})X_{t-1} + a_2(X_{t-2})X_{t-2} + \sigma\varepsilon_t \tag{5.3}$$

to the first 102 observations, reserving the last 12 observations to evaluate the prediction. The $a_1(\cdot)$ and $a_2(\cdot)$ in (5.3) are unknown functions which are estimated by using locally linear smoothers. Fan and Yao ([14], p. 327) plot the estimates $\hat{a}_1(\cdot)$ and $\hat{a}_2(\cdot)$, which are approximately constant for $X_{t-2} < 2.7$ with $\hat{a}_1(X_{t-2}) \approx 1.3$ and $\hat{a}_2(X_{t-2}) \approx -0.2$, and which are approximately linear for $X_{t-2} \geq 2.7$. For comparison, Fan and Yao [14] also fit the following SETAR(2) model to the same set of data:

$$\widehat{X}_t = \begin{cases} 0.424 + 1.255X_{t-1} - 0.348X_{t-2}, & X_{t-2} \leq 2.981, \\ 1.882 + 1.516X_{t-1} - 1.126X_{t-2}, & X_{t-2} > 2.981. \end{cases} \tag{5.4}$$

Because of the close resemblance of the fitted SETAR(2) and FAR(2,2), they share certain ecological interpretations. In particular, the difference of the fitted coefficients in each regime can be explained by using the phase dependence and the density dependence in the predator-prey structure. The phase dependence refers to the difference in the behavior of preys (snowshoe hare) and predators (lynx) in hunting and escaping at the decreasing and increasing phases of population dynamics, while the density dependence is the relationship between the reproduction rates of the animals and their abundance. More discussion on these ecological interpretations can be found in [42].

To evaluate the predictions of FAR (2,2), Fan and Yao ([14], p. 324) use the one-step ahead forecast (denoted by $\widehat{X}_t$) and the iterative two-step-ahead forecast (denoted by $\tilde{X}_t$), which are defined by

$$\widehat{X}_t := \hat{a}_1(X_{t-2})X_{t-1} + \hat{a}_2(X_{t-2})X_{t-2}, \quad \tilde{X}_t := \hat{a}_1(X_{t-2})\widehat{X}_{t-1} + \hat{a}_2(X_{t-2})X_{t-2}.$$

The predictions of SETAR(2) are similarly defined. The out-sample prediction absolute errors ($|\hat{X}_t - X_t|$ and $|\tilde{X}_t - X_t|$) of the last 12 observations are reported in Table 1. Based on the average of these 12 absolute prediction errors (AAPE), FAR(2,2) performs slightly better than SETAR(2). Other nonparametric time series models for the Canadian lynx data include the projection pursuit regression (PPR) model fitted by Lin and Pourahmadi [36] who found that SETAR outperforms PPR in terms of one-step-ahead forecasts, and neural network models which Kajitani, Hipel and McLeod [25] found to be "just as good or better than SETAR models for one-step out-of-sample forecasting of the lynx data."

A substantive approach is adopted by Royama ([41], Chapter 5). Instead of building the statistical model first and using ecology to interpret the fitted model later, Royama starts with ecological mechanisms and population dynamics. Letting $R_t = X_{t+1} - X_t$ denote the log reproductive rate from year t to $t+1$, he considers nonlinear dynamics of the form $R_t = f(X_t, \ldots, X_{t-h+1}) + u_t$, where u_t is a zero-mean random disturbance, and emphasizes that "our ultimate goal is to determine the reproduction surface f and to find an appropriate model which reasonably approximates to it," with f satisfying the following two conditions in view of ecological considerations: There exists X^* such that $f(X^*, \ldots, X^*) = 0$, and R_t has to be bounded above because "no animal can produce infinite number of offspring"

TABLE 1
Absolute prediction errors of one-step-ahead (1 yr) and iterative two-step-ahead (2 yr) forecasts and their 12-year average (AAPE).

		Model (5.3) FAR(2,2)		Model (5.4) SETAR(2)		Model (5.6) Logistic		Model (5.8a) Logistic-MARS	
Year	X_t	1 yr	2 yr	1 yr	2 yr	1 yr	2 yr	1 yr	2 yr
1923	3.054	0.157	0.156	0.187	0.090	0.178	0.075	0.188	0.082
1924	3.386	0.012	0.227	0.035	0.269	0.077	0.281	0.057	0.286
1925	3.553	0.021	0.035	0.014	0.038	0.057	0.153	0.073	0.120
1926	3.468	0.008	0.037	0.022	0.000	0.012	0.077	0.023	0.140
1927	3.187	0.085	0.101	0.059	0.092	0.020	0.018	0.122	0.168
1928	2.723	0.055	0.086	0.075	0.015	0.128	0.098	0.002	0.159
1929	2.686	0.135	0.061	0.273	0.160	0.179	0.004	0.009	0.012
1930	2.821	0.016	0.150	0.026	0.316	0.004	0.216	0.010	0.001
1931	3.000	0.017	0.037	0.030	0.062	0.005	0.010	0.013	0.025
1932	3.201	0.007	0.014	0.060	0.043	0.048	0.042	0.021	0.005
1933	3.424	0.089	0.098	0.076	0.067	0.124	0.184	0.066	0.091
1934	3.531	0.053	0.175	0.072	0.187	0.083	0.245	0.011	0.087
AAPE		0.055	0.095	0.073	0.112	0.075	0.117	0.050	0.098

(see [41], p. 50, 154, 178). In Chapter 4 of [42], Royama introduces the (first-order) logistic model of $f(X_t) = r_m - \exp\{-a_0 - a_1 X_{t-1}\}$ to incorporate competition over an available resource. Here r_m is the maximum biologically realizable reproduction rate, i.e. $R_t \leq r_m$ for all t; see [42], Section 4.2.5. An implicit assumption of the model is that the resource being depleted during a time step will be recovered to the same level by the onset of the next time step. This assumption can be relaxed if a linear combination of $X_{t-j} (j = 1, \ldots, h)$ with $h > 1$ is used in the exponential term of f, yielding a higher-order logistic model; see [41], p. 153.

Chapter 5 of Royama [41] examines the autocorrelation function and the partial autocorrelation function of the Canadian lynx series and concludes that h should be set to 2, which corresponds to the model

$$X_t - X_{t-1} = r_m - \exp\{-a_0 - a_1 X_{t-1} - a_2 X_{t-2}\} + u_{t-1}, \tag{5.5}$$

where r_m, a_0, a_1 and a_2 are unknown parameters that need to be estimated; see [41], p. 190-191. From the scatterplot of $R_{t-1} = X_t - X_{t-1}$ versus X_{t-2}, Royama guesses $r_m \approx 0.6$ and $X^* \approx 3$. He uses this together with trial and error to obtain the estimate $(\widehat{r}_m, \widehat{a}_0, \widehat{a}_1, \widehat{a}_2) = (0.597, 2.526, 0.838, -1.508)$, but finds that the asymmetric cycle of the fitted model does not match the observed cycle from the data well. Moreover, the fitted autocorrelation function decays too fast in comparison with the sample autocorrelation function.

Instead of his *ad hoc* estimates, we can use nonlinear least squares, initialized at his estimates, to estimate the parameters of (5.5), yielding

$$X_t - X_{t-1} = 0.460 - \exp\{-3.887 - 0.662 X_{t-1} + 1.663 X_{t-2}\} + u_{t-1}, \tag{5.6}$$

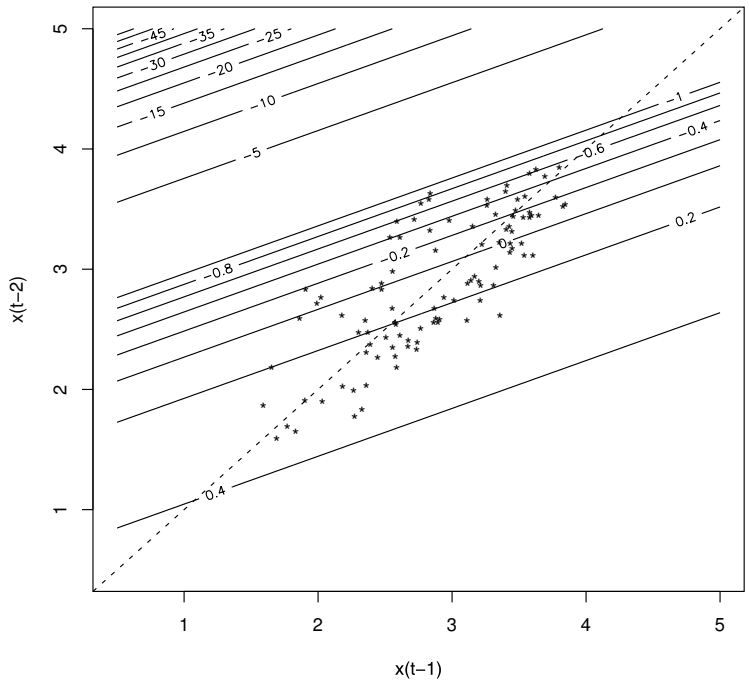


FIG 2. *Contour plot of* $\hat{R}_{t-1} = \widehat{X_t - X_{t-1}}$ *of the logistic model (5.6). The observations are marked by* $*$. *The dotted line is* $X_{t-2} = X_{t-1}$. *The intersection of this line and the contour numbered 0 gives the equilibrium* X^*.

which implies that the maximum logarithmic reproduction rate is 0.460, i.e., the population can grow at most $10^{0.46} = 2.884$ times per year. Figure 2, top left corner, shows a negative contour of the response surface of the fitted model (5.6). This implies that the population size can drop sharply in the region $X_{t-2} > 3.5$ and $X_{t-1} < 2.5$, leading to extinction in the upper left part of this region. Whereas (5.6) does not rule out the possibility of X_t diverging to $-\infty$, extinction occurs as soon as X_t falls below 0 (or equivalently, the population size 10^{X_t} falls below 1).

Note that one can also derive bounds on the logarithmic reproduction rates from the empirical approach. Figure 3 is the plot of the limit cycle generated by the skeleton of the fitted model (5.4). The limit cycle is of period 8 years. The maximum and the minimum logarithmic reproduction rates, attained at years 1 and 5 in Figure 3, are 0.212 and -0.269, respectively. That is, the population grows at most $10^{0.212} = 1.629$ times per year and diminishes by at most a factor of $10^{-0.269}$=0.538 per year. Moreover, the limit cycle of (5.4) implies an infinite loop of expansion and contraction and rules out the possibility of extinction. These are consequences of adopting an empirical approach because the data are distributed along the main diagonal of Figure 2, but not its top left corner nor its lower right corner. In order to deduce the behavior of the reproduction rates in these regions, mechanistic modeling is essential. On the other hand, the empirical approach uses the observed data better and gives more accurate forecasts. Table 1 compares the prediction performance of FAR(2,2) and SETAR(2) with that of the logistic model (5.5). The fitted logistic model provides the worst AAPE of one-step-ahead and iterative two-step-ahead forecasts. Moreover, instead of characterizing the equilibrium with limit cycles, the logistic model only gives two equilibrium points, with one corresponding to extinction and the other equal to $X^* = \{a_0 + \log(r_m)\}/(a_1 + a_2) = 3.107$ (the intersection of the line $X_{t-1} = X_{t-2}$ and the contour of $f = 0$ in Figure 2.)

We next apply the combined substantive-empirical approach of Section 4 to these data, using the substantive model (5.5) to provide one of the basis functions in the

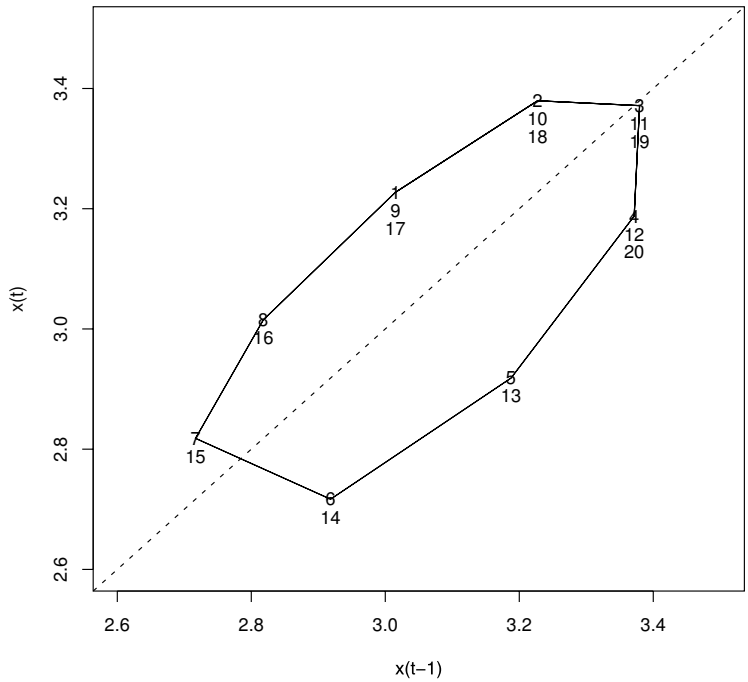


FIG 3. *Limit cycle of the skeleton of the SETAR(2) model (5.4). The dotted line is* $X_t = X_{t-1}$.

semiparametric model

$$X_t - X_{t-1} = r_m - \exp\{-a_0 - a_1 X_{t-1} - a_2 X_{t-2}\} + g(X_{t-1}, X_{t-2}) I\{(X_{t-1}, X_{t-2}) \in S\} + u_{t-1}, \tag{5.7}$$

where g is an unknown function and S is a region containing the observations that will be specified later. Moreover, the difference equation (5.7) has the boundary constraint $X_{t-1} + r_m - \exp\{-a_0 - a_1 X_{t-1} - a_2 X_{t-2}\} + g(X_{t-1}, X_{t-2}) I\{(X_{t-1}, X_{t-2}) \in S\} \geq 0$. The lynx population becomes extinct as soon as this boundary condition is violated. Model (5.7) can be fitted by using the backfitting algorithm. Specifically, model (5.5) is estimated first and then the residuals are used as the response variable in nonparametric regression on the predictor variable (X_{t-1}, X_{t-2}). The difference between the observed $X_t - X_{t-1}$ and the fitted g is then used as the response variable in (5.5), whose parameters can be estimated by nonlinear least squares. The algorithm of multivariate adaptive regression splines (MARS) developed by Friedman (1991) is used for estimating g for the first step in each iteration of the above backfitting procedure (the function "mars" in the package of "mda" in R can be used). This kind of iteration sheme has been used in fitting *partly linear* models, where the parametric component is a linear regression model and the nonparametric component is often fitted by using kernel regression; see [8, 13, 20]. The fitted response surface is

$$X_t - X_{t-1} = 1.319 - \exp\{-0.224 - 0.205 X_{t-1} + 0.343 X_{t-2}\} + \hat{g}(X_{t-1}, X_{t-2}) I\{(X_{t-1}, X_{t-2}) \in S\} + u_{t-1}, \tag{5.8a}$$

$$\hat{g}(X_{t-1}, X_{t-2}) = 2.294(X_{t-1} - 3.224)_+ (X_{t-2} - 2.864)_+ - 1.572(X_{t-1} - 3.202)_+ - 0.851(X_{t-2} - 3.202)_+. \tag{5.8b}$$

We evaluate this fitted model by using the out-sample prediction criterion. Table 1 shows that (5.8a) gives the smallest AAPE for one-step-ahead forecasts among all models considered, and that the AAPE for iterative two-step-ahead forecasts of (5.8a) is comparable to the smallest one provided by FAR(2,2). The region S in (5.8a) is chosen to be the oblique rectangle whose edges are defined by the sample means ± 3 standard deviations of the principal components of the bivariate sample of (X_{t-1}, X_{t-2}); see Figure 4 which shows that this region contains not only the in-sample data but also the out-sample data. Figure 5 gives the contour plot of the fitted model (5.8a). The logarithmic growth rate at its top left corner is about -2, which shows a strong possibility of extinction even though the magnitude is less drastic than that in Figure 2 for (5.6). The inclusion of tensor products of univariate splines in (5.8a) would have produced positive probability limits of X_t diverging to ∞ or to $-\infty$ if (X_{t-1}, X_{t-2}) had not been confined to a compact region. On the other hand, with an absorbing barrier at 0 and with (5.8b) only applicable inside the compact set S, Markov chains of the type (5.8a) not only have stationary distributions but are also geometrically ergodic under mild assumptions on the random disturbances u_t (e.g., to ensure irreducibity); see [39].

6. Conclusion

In his concluding remarks, Cox [11] noted that for successful use of statistical models in particular applications, "large elements of subject-matter judgment and technical statistical expertise are usually essential. Indeed, it is precisely the need for this combination that makes our subject such an interesting and demanding one." We

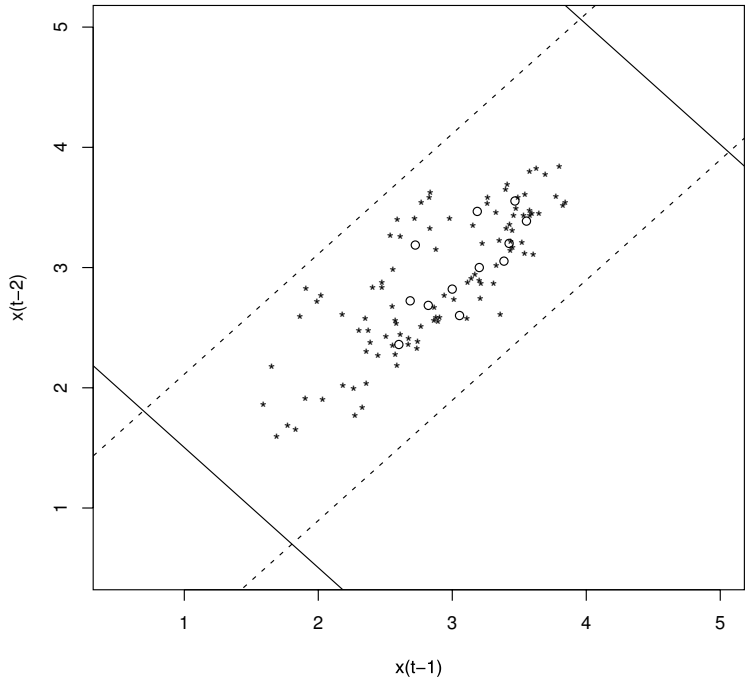


FIG 4. *The oblique rectangle S formed by ± 3 standard deviations away from the sample means of the principal components of (X_{t-1}, X_{t-2}). The in-sample and out-sample observations are marked by $*$ and o, respectively.*

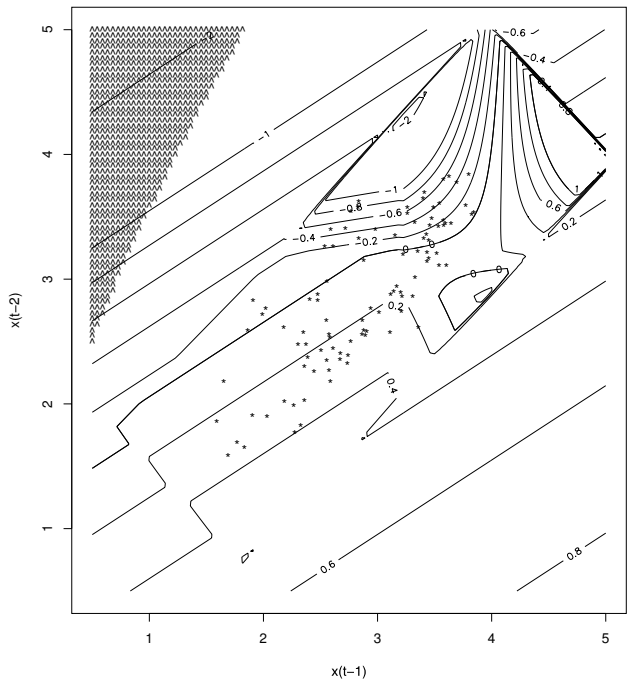


FIG 5. *Contour plot of $\hat{R}_{t-1} = X_t - \widehat{X_{t-1}}$ of the logistic-MARS model (5.7). The observations are marked by $*$. The shaded region corresponds to extinction.*

have followed up on his remarks here with a combined subject-matter and statistical modeling approach to time series analysis, which we illustrate for the "particular applications" of option pricing and population dynamics of the Canadian lynx. In particular, for the Canadian lynx data, we have shown how statistical modeling for data-rich regions of (X_{t-1}, X_{t-2}) can be combined effectively with "subject-matter judgment" which is the only reliable guide for sparse-data regions.

Acknowledgments

Lai's research was supported by the National Science Foundation grant DMS-0305749. Wong's research was supported by the Research Grants Council of Hong Kong under grant CUHK6158/02E.

References

[1] AitSahlia, F. and Lai, T. L. (2001). Exercise boundaries and efficient approximations to American option prices and hedge parameters. *J. Comput. Finance* **4** 85–103.

[2] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoid function. *IEEE Trans. Information Theory* **39** 930–945.

[3] Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *J. Political Economy* **81** 637–659.

[4] Broadie, M. and Detemple, J. (1996). American option valuation: New bounds, approximations, and a comparison of existing methods. *Rev. Financial Studies* **9** 1121–1250.

[5] Broadie, M., Detemple, J., Ghysels, E. and Torres, O. (2000). Nonparametric estimation of American options' exercise boundaries and call prices. *J. Econ. Dynamics & Control* **24** 1829–1857.

[6] Campbell, M. J. and Walker, A.M. (1977). A survey of statistical work on the McKenzie River series of annual Canadian lynx trappings for the years 1821-1934, and a new analysis. *J. Roy. Statist. Soc. Ser. A* **140** 411–431.

[7] Carr, P., Jarrow, R. and Myneni, R. (1992). Alternative characterizations of American put options. *Math. Finance* **2** 87–106.

[8] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–146.

[9] Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88** 298–308.

[10] Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88** 955–967.

[11] Cox, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5** 169–174.

[12] Cox, J., Ross, S. and Rubinstein, M. (1979). Option pricing: A simplified approach. *J. Financial Econ.* **7** 229–263.

[13] Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

[14] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series*. Springer-Verlag, New York.

[15] Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–142.

[16] Goodwin, G. C., Ramadge, P. J. and Caines, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* **19** 829–853.
[17] Guo, L. and Chen, H. F. (1991). The Åström-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers. *IEEE Trans. Automat. Contr.* **36** 802–812.
[18] Haggan, V. and Ozaki, T. (1981). Modelling non-linear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68** 189–196.
[19] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.
[20] Heckman, N. E. (1988). Minimax estimates in a semiparametric model. *J. Amer. Statist. Assoc.* **83** 1090–1096.
[21] Hull, J. C. (2006). *Options, Futures and Other Derivatives*, 6th edn. Pearson Prentice Hall, Upper Saddle River, NJ.
[22] Hutchinson, J. M., Lo, A. W. and Poggio, T. (1994). A nonparameric approach to pricing and hedging derivative securities via learning networks. *J. Finance* **49** 851–889.
[23] Jacka, S. D. (1991). Optimal stopping and the American put. *Math. Finance* **1** 1–14.
[24] Ju, N. (1998). Pricing an American option by approximating its early exercise boundary as a multipiece exponential function. *Rev. Financial Studies* **11** 627–646.
[25] Kajitani, Y., Hipelm, K. W. and McLeod, A. I. (2005). Forecasting nonlinear time series with feed-forward neural networks: A case study of Canadian Lynx data. *J. Forecasting* **24** 105–117.
[26] Lai, T. L. and Lim, T. W. (2006). A new approach to pricing and hedging options with transaction costs. Tech. Report, Dept. Statistics, Stanford Univ.
[27] Lai, T. L. and Wei, C. Z. (1987). Asymptotically efficient self-tuning regulators. *SIAM J. Control Optim.* **25** 466–481.
[28] Lai, T. L. and Wong, S. P. (2001). Stochastic neural networks with applications to nonlinear time series. *J. Amer. Statist. Assoc.* **96** 968–981.
[29] Lai, T. L. and Wong, S. P. (2004). Valuation of American options via basis functions. *IEEE Trans. Automat. Contr.* **49** 374–385.
[30] Lai, T. L. and Ying, Z. (1991). Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems. *SIAM J. Control Optim.* **29** 1091–1127.
[31] Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statist. Sci.* **5** 160–168.
[32] Lewis, P. A. W. and Ray, B. K. (1993). Nonlinear modeling of multivariate and categorical time series using multivariate adaptive regression splines. In *Dimension Estimation and Models* (H. Tong, ed). World Sci. Publishing, River Edge, NJ, pp. 136–169.
[33] Lewis, P. A. W. and Ray, B. K. (2002). Nonlinear modelling of periodic threshold autoregressions using TSMARS. *J. Time Ser. Anal.* **23** 459–471.
[34] Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *J. Amer. Statist. Assoc.* **86** 864–877.
[35] Lim, K. S. (1987). A comparative study of various univariate time series models for Canadian lynx data. *J. Time Ser. Anal.* **8** 161–176.

[36] Lin, T. C. and Pourahmadi, M. (1998). Nonparametric and nonlinear models and data mining in time series: a case-study on the Canadian lynx data. *Appl. Statist.* **47** 187–201.

[37] Merton, R. C. (1973). Theory of rational option pricing. *Bell J. Econ. & Management Sci.* **4** 141–181.

[38] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York.

[39] Moran, P. A. P. (1953). The statistical analysis of the Canadian lynx cycle, I: Structure and prediction. *Austral. J. Zoology* **1** 163–173.

[40] Ross, S. A. (1987). Finance. In *The New Palgrave: A Dictionary of Economics* (J. Eatwell, M. Milgate and P. Newman, eds.), Vol. 2. Stockton Press, New York, pp. 322–336.

[41] Royama, T. (1992). *Analytical Population Dynamics*. Chapman & Hall, London.

[42] Stenseth, N. C., Chan, K. S., Tong, H., Boonstra, R., Boutin, S., Krebs, C. J.,Post, E., O'Donoghue, M., Yoccoz, N. G., Forchhammer, M. C. and Hurrell, J. W. (1998). From patterns to processes: Phase and density dependencies in the Canadian lynx cycle. *Proc. Natl. Acad. Sci. USA* **95** 15430–15435.

[43] Tong, H. (1977). Some comments on the Canadian lynx data. *J. Roy. Statist. Soc. Ser. A* **140** 432–435.

[44] Tong, H. (1990). *Nonlinear Time Series.* Oxford University Press, Oxford.

[45] Tong H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data (with Discussion). *J. Roy. Statist. Soc. Ser. B* **42** 245–292.

[46] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM Press, Philadelphia.

[47] Weigend, A. and Gershenfeld, N. (1993). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, Reading, MA.

[48] Weigend, A., Rumelhart, D. and Huberman, B. (1991). Predicting Sunspots and Exchange Rates with Connectionist Networks. In *Nonlinear Modeling and Forecasting* (Casdagli, M. and Eubank, S., eds.). Addison Wesley, Redwood City, CA, 395–432.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 210–222

DOI: 10.1214/074921706000001058

Multivariate volatility models

Ruey S. Tsay[1]

University of Chicago

Abstract: Correlations between asset returns are important in many financial applications. In recent years, multivariate volatility models have been used to describe the time-varying feature of the correlations. However, the curse of dimensionality quickly becomes an issue as the number of correlations is $k(k-1)/2$ for k assets. In this paper, we review some of the commonly used models for multivariate volatility and propose a simple approach that is parsimonious and satisfies the positive definite constraints of the time-varying correlation matrix. Real examples are used to demonstrate the proposed model.

1. Introduction

Let $r_t = (r_{1t}, \ldots, r_{kt})'$ be a vector of returns (or log returns) of k assets at time index t. Let F_{t-1} be the sigma field generated by the past information at time index $t-1$. We partition the return r_t as

$$r_t = \mu_t + e_t, \tag{1}$$

where $\mu_t = E(r_t|F_{t-1})$ is the conditional mean of the return given F_{t-1} and e_t is the innovation (or noise term) satisfying $e_t = \Sigma_t^{1/2}\epsilon_t$ such that

$$\mathrm{Cov}(e_t|F_{t-1}) = \mathrm{Cov}(r_t|F_{t-1}) = \Sigma_t, \tag{2}$$

where $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{kt})'$ is a sequence of independently and identically distributed random vectors with mean zero and identity covariance matrix, and $\Sigma_t^{1/2}$ is the symmetric square-root matrix of a positive-definite covariance matrix Σ_t, that is, $\Sigma_t^{1/2}\Sigma_t^{1/2} = \Sigma_t$. In the literature, Σ_t is often referred to as the volatility matrix. Volatility modeling is concerned with studying the evolution of the volatility matrix over time. For asset returns, behavior of the conditional mean μ_t is relatively simple. In most cases, μ_t is simply a constant. In some cases, it may assume a simple vector autoregressive model. The volatility matrix Σ_t, on the other hand, is much harder to model, and most GARCH studies in the literature focus entirely on modeling Σ_t.

The conditional covariance matrix Σ_t can be written as

$$\Sigma_t = D_t R_t D_t \tag{3}$$

where D_t is a diagonal matrix consisting of the conditional standard deviations of the returns, i.e., $D_t = \mathrm{diag}\{\sqrt{\sigma_{11,t}}, \ldots, \sqrt{\sigma_{kk,t}}\}$ with $\sigma_{ij,t}$ being the (i,j)th element of Σ_t, and R_t is the correlation matrix.

[1]Graduate School of Business, University of Chicago, 5807 S. Woodlawn Avenue, Chicago, IL 60637, e-mail: ruey.tsay@ChicagoGSB.edu

AMS 2000 subject classifications: primary 62M10; secondary 62M20.

Keywords and phrases: multivariate GARCH model, BEKK model, positive definite matrix, volatility series.

In recent years, many studies extend the univariate generalized autoregressive conditional heteroscedastic (GARCH) model of Bollerslev [2] to the multivariate case for modeling the volatility of multiple asset returns; see the recent article [1] for a survey. Multivariate volatility models have many important applications in finance and statistics. They can be used to study the correlations between asset returns. These correlations play an important role in asset allocation, risk management, and portfolio selection. There are two major difficulties facing the generalization, however. First of all, the dimension of volatility matrix increases rapidly as the number of asset increases. Indeed, there are $k(k+1)/2$ variances and covariances for k asset returns. Second, for asset returns the covariance matrix is time-varying and positive definite. Many of the multivariate volatility models proposed in the literature fail to satisfy the positive-definite constraints, e.g., the diagonal VEC model [3], even though they are easy to understand and apply.

The goal of this paper is to propose a simple approach to modeling multivariate volatility. The proposed model is kept parsimonious in parameterization to overcome the difficulty of curse of dimensionality. In addition, a simple structure equation is imposed to ensure that the resulting time-varying covariance matrices are positive definite. On the other hand, the proposed model is not very flexible and may encounter lack of fit when the dimension is high. To safe guard against model inadequacy, we consider model checking using some bootstrap methods to generate finite-sample critical values of the test statistics used.

The paper is organized as follows. In Section 2, we briefly review some of the multivariate volatility models relevant to the proposed model. Section 3 considers the proposed model whereas Section 4 contains applications to daily returns of foreign exchange rates and U.S. stocks. Section 5 concludes.

2. A brief review of vector volatility models

Many multivariate volatility models are available in the literature. In this section, we briefly review some of those models that are relevant to the proposed model. We shall focus on the simple models of order $(1,1)$ in our discussion because such models are typically used in applications and the generalization to higher-order models is straightforward. In what follows, let a_{ij} denote the (i,j)th element of the matrix A and u_{it} be the ith element of the vector u_t.

VEC model. For a symmetric $n \times n$ matrix A, let $\text{vech}(A)$ be the half-stacking vector of A, that is, $\text{vech}(A)$ is a $n(n+1)/2 \times 1$ vector obtained by stacking the lower triangular portion of the matrix A. Let $h_t = \text{vech}(\Sigma_t)$ and $\eta_t = \text{vech}(e_t e_t')$. Using the idea of exponential smoothing, Bollerslev et al. [3] propose the VEC model

$$h_t = c + A\eta_{t-1} + Bh_{t-1} \tag{4}$$

where c is a $k(k+1)/2$-dimensional vector, and A and B are $k(k+1)/2 \times k(k+1)/2$ matrices. This model contains several weaknesses. First, the model contains $k(k+1)[k(k+1)+1]/2$ parameters, which is large even for a small k. For instance, if $k=3$, then the model contains 78 parameters, making it hard to apply in practice. To overcome this difficulty, Bollerslev et al. [3] further suggest that both A and B matrices of Eq. (4) are constrained to be diagonal. The second weakness of the model is that the resulting volatility matrix Σ_t may not be positive definite.

BEKK model. A simple BEKK model of Engle and Kroner [5] assumes the form

$$\Sigma_t = C'C + A'e_{t-1}e'_{t-1}A + B'\Sigma_{t-1}B \tag{5}$$

where C, A, and B are $k \times k$ matrices but C is upper triangular. An advantage of the BEKK model is that Σ_t is positive definite if the diagonal elements of C is positive. On the other hand, the model contains many parameters that do not represent directly the impact of e_{t-1} or Σ_{t-1} on the elements of Σ_t. In other words, it is hard to interpret the parameters of a BEKK model. Limited experience also shows that many parameter estimates of the BEKK model in Eq. (5) are statistically insignificant, implying that the model is overparameterized.

Using the standardization of Eq. (3), one can divide the multivariate volatility modeling into two steps. The first step is to specify models for elements of the D_t matrix, and the second step is to model the correlation matrix R_t. Two such approaches have been proposed in the literature. In both cases, the elements $\sigma_{ii,t}$ are assumed to follow a univariate GARCH model. In other words, $\sigma_{ii,t}$ are based entirely on the i-the return series.

Dynamic correlation model of Tse and Tsui. In [8], the authors propose that (a) the individual volatility $\sigma_{ii,t}$ can assume any univariate GARCH models, and (b) the correlation matrix R_t of Eq. (3) follows the model

$$R_t = (1 - \lambda_1 - \lambda_2)R + \lambda_1 \Psi_{t-1} + \lambda_2 R_{t-1} \tag{6}$$

where λ_1 and λ_2 are non-negative parameters satisfying $0 \leq \lambda_1 + \lambda_2 < 1$, R is a $k \times k$ positive definite parameter matrix with $R_{ii} = 1$ and Ψ_{t-1} is the $k \times k$ correlation matrix of some recent asset returns. For instance, if the most recent m returns are used to define Ψ_{t-1}, then the (i, j)th element of Ψ_{t-1} is given by

$$\psi_{ij,t-1} = \frac{\sum_{v=1}^{m} u_{i,t-v}u_{j,t-v}}{\sqrt{(\sum_{v=1}^{m} u_{i,t-v}^2)(\sum_{v=1}^{m} u_{j,t-v}^2)}},$$

where $u_{it} = e_{it}/\sqrt{\sigma_{ii,t}}$. If $m > k$, then Ψ_{t-1} is positive definite almost surely. This in turn implies that R_t is positive definite almost surely. We refer to this model as a $\text{DCC}_T(m)$ model. In practice, one can use the sample correlation matrix of the data to estimate R in order to simplify the calculation. Indeed, this is the approach we shall take in this paper.

From the definition, the use of $\text{DCC}_T(m)$ model involves two steps. In the first step, univariate GARCH models are built for each return series. At step 2, the correlation matrix R_t of Eq. (6) is estimated across all return series via the maximum likelihood method. An advantage of the $\text{DCC}_T(m)$ model is that the resulting correlation matrices are positive definite almost surely. In addition, the model is parsimonious in parameterization because the evolution of correlation matrices is governed by two parameters. On the other hand, strong limitation is imposed on the time evolution of the correlation matrices. In addition, it is hard to interpret the results of the two-step estimation. For instance, it is not clear what is the joint distribution of the innovation e_t of the return series.

Dynamic correlation model of Engle. A similar correlation model is proposed by Engle [4]. Here the correlation matrix R_t follows the model

$$R_t = W_t^{-1} Q_t W_t^{-1} \tag{7}$$

where $Q_t = [q_{ij,t}]$ is a positive-definite matrix, $W_t = \text{diag}\{\sqrt{q_{11,t}}, \ldots, \sqrt{q_{kk,t}}\}$ is a normalization matrix, and the elements of Q_t are given by

$$Q_t = (1 - \alpha_1 - \alpha_2)\bar{Q} + \alpha_1 u_{t-1} u'_{t-1} + \alpha_2 Q_{t-1},$$

where u_t is the standardized innovation vector with elements $u_{it} = e_{it}/\sqrt{\sigma_{ii,t}}$, $\bar{Q}$ is the sample covariance matrix of u_t, and α_1 and α_2 are non-negative scalar parameters satisfying $0 < \alpha_1 + \alpha_2 < 1$. We refer to this model as the DCC_E model.

Compared with the $\text{DCC}_T(m)$ model, the DCC_E model only uses the most recent standardized innovation to update the time-evolution of the correlation matrix. Since $u_{t-1}u'_{t-1}$ is singular for $k > 1$ and is, in general, not a correlation matrix, and the matrix Q_t must be normalized in Eq. (7) to ensure that R_t is indeed a correlation matrix. Because a single innovation is more variable than the correlation matrix of m standardized innovations, the correlations of a DCC_E model tend to be more variable than those of a $\text{DCC}_T(m)$ model.

To better understand the difference between $\text{DCC}_T(m)$ and DCC_E models, consider the correlation $\rho_{12,t}$ of the first two returns in r_t. For $\text{DCC}_T(m)$ model,

$$\rho_{12,t} = (1 - \lambda_1 - \lambda_2)\rho_{12} + \lambda_2 \rho_{12,t-1} + \lambda_1 \frac{\sum_{v=1}^{m} u_{1,t-v} u_{2,t-v}}{\sqrt{(\sum_{v=1}^{m} u_{1,t-v}^2)(\sum_{v=1}^{m} u_{2,t-v}^2)}}.$$

On the other hand, for the DCC_E model,

$$\rho_{12,t} = \frac{\alpha^* \bar{q}_{12} + \alpha_1 u_{1,t-1} u_{2,t-1} + \alpha_2 q_{12,t-1}}{\sqrt{(\alpha^* \bar{q}_{11} + \alpha_1 u_{1,t-1}^2 + \alpha_2 q_{11,t-1})(\alpha^* \bar{q}_{22} + \alpha_1 u_{2,t-1}^2 + \alpha_2 q_{22,t-1})}},$$

where $\alpha^* = 1 - \alpha_1 - \alpha_2$. The difference is clearly seen.

3. Proposed models

We start with the simple case in which the effects of positive and negative past returns on the volatility are symmetric. The case of asymmetric effects is given later.

3.1. Multivariate GARCH models

In this paper, we propose the following model

$$r_t = \mu_t + e_t, \quad \mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i}, \quad e_t = \Sigma^{1/2} \epsilon_t \tag{8}$$

where p is a non-negative integer and $\{\epsilon_t\}$ is a sequence of independent and identically distributed multivariate Student-t distribution with v degrees of freedom. The probability density function of ϵ_t is

$$f(\epsilon) = \frac{\Gamma((v+k)/2)}{[\pi(v-2)]^{k/2}\Gamma(v/2)} [1 + (v-2)^{-1}\epsilon'\epsilon]^{-(v+k)/2}.$$

The variance of each component of ϵ_t is 1. The volatility matrix is standardized as Eq. (3) with elements of D_t and the correlation matrix R_t satisfying

$$D_t^2 = \Lambda_0 + \Lambda_1 D_{t-1}^2 + \Lambda_2 G_{t-1}^2, \tag{9}$$

$$R_t = (1 - \theta_1 - \theta_2)\bar{R} + \theta_1 \psi_{t-1} + \theta_2 R_{t-1}, \tag{10}$$

where $G_t = \text{diag}\{e_{1t}, \ldots, e_{kt}\}$, $\Lambda_i = \text{diag}\{\ell_{11,i}, \ldots, \ell_{kk,i}\}$ are diagonal matrices such that $\ell_{ii,1} + \ell_{ii,2} < 1$ and $0 \leq \ell_{ii,j}$ for $i = 1, \ldots, k$ and $j = 1, 2$, $\bar{R}$ is the sample correlation matrix, θ_i are non-negative real numbers satisfying $\theta_1 + \theta_2 < 1$, and ψ_{t-1} is the sample correlation matrix of the last m innovations as defined in the $\text{DCC}_T(m)$ model of Eq. (6). We use $m = k + 2$ in empirical data analysis.

This model uses univariate GARCH(1,1) models for the conditional variance of components of r_t and a combination of the correlation matrix equations of the $\text{DCC}_T(m)$ and DCC_E models for the correlation. The order of GARCH models can be increased if necessary, but we use (1,1) for simplicity. In addition, Λ_1 and Λ_2 can be generalized to non-diagonal matrices. However, we shall keep the simple structure in Eq. (9) and (10) for ease in application and interpretation.

The proposed model differs from the $\text{DCC}_T(m)$ model in several ways. First, the proposed model uses a multivariate Student-t distribution for innovation so that the degrees of freedom are the same for all asset returns. This simplifies the model interpretation at the expense of decreased flexibility. Second, the proposed model uses sample correlation matrix $\bar{R}$ to reduce the number of parameters. Third, the proposed model uses joint estimation whereas the $\text{DCC}_T(m)$ model performs separate estimations for variances and correlations.

3.2. Model with leverage effects

In financial applications, large positive and negative shocks to an asset have different impacts on the subsequent price movement. In volatility modeling, it is expected that a large negative shock would lead to increased volatility as a big drop in asset price is typically associated with bad news which, in turn, means higher risk for the investment. This phenomenon is referred to as the *leverage* effect in volatility modeling. The symmetry of GARCH model in Eq. (9) keeps the model simple, but fails to address the leverage effect. To overcome this shortcoming, we consider the modified model

$$D_t^2 = \Lambda_0 + \Lambda_1 D_{t-1}^2 + \Lambda_2 G_{t-1}^2 + \Lambda_3 L_{t-1}^2, \tag{11}$$

where Λ_i $(i = 0, 1, 2)$ are defined as before, Λ_3 is a $k \times k$ diagonal matrix with non-negative diagonal elements, and L_{t-1} is also a $k \times k$ diagonal matrix with diagonal elements

$$L_{ii,t-1} = \begin{cases} e_{i,t-1} \text{ if } e_{i,t-1} < 0, \\ 0 \qquad \text{otherwise.} \end{cases}$$

In Eq. (11), we assume that $0 < \sum_{j=1}^{3} \ell_{ii,j} \leq 1$ for $i = 1, \ldots, k$. This is a sufficient condition for the existence of volatility.

From the definition, a positive shock $e_{i,t-1}$ affects the volatility via $\ell_{ii,2} e_{i,t-1}^2$. A negative shock, on the other hand, contributes $(\ell_{ii,2} + \ell_{ii,3}) e_{i,t-1}^2$ to the volatility. Checking the significance of $\ell_{ii,3}$ enables us to draw inference on the leverage effect.

4. Application

We illustrate the proposed model by considering some daily asset returns. First, we consider a four-dimensional process consisting of two equity returns and two exchange rate returns. Second, we consider a 10-dimensional equity returns. In both examples, we use $m = k + 2$ to estimate the local correlation matrices ψ_{t-1} in Eq. (10).

Example 1. In this example, we consider the daily exchange rates between U.S. Dollar versus European Euro and Japanese Yen and the stock prices of IBM and Dell from January 1999 to December 2004. The exchange rates are the noon spot rate obtained from the Federal Reserve Bank of St. Louis and the stock returns are from the Center for Research in Security Prices (CRSP). We compute the simple returns of the exchange rates and remove returns for those days when one of the markets was not open. This results in a four-dimensional return series with 1496 observations. The return vector is $r_t = (r_{1t}, r_{2t}, r_{3t}, r_{4t})'$ with r_{1t} and r_{2t} being the returns of Euro and Yen exchange rate, respectively, and r_{3t} and r_{4t} are the returns of IBM and Dell stock, respectively. All returns are in percentages. Figure 1 shows the time plot of the return series. From the plot, equity returns have higher variability than the exchange rate returns, and the variability of equity returns appears to be decreasing in recent years. Table 1 provides some descriptive statistics of the return series. As expected, the means of the return are essentially zero and all four series have heavy tails with positive excess kurtosis.

The equity returns have some serial correlations, but the magnitude is small. If multivariate Ljung-Box statistics are used, we have $Q(3) = 59.12$ with p-value 0.13 and $Q(5) = 106.44$ with p-value 0.03. For simplicity, we use the sample mean as the mean equation and apply the proposed multivariate volatility model to the mean-corrected data. In estimation, we start with a general model, but add some equality constraints as some estimates appear to be close to each other. The results are given in Table 2 along with the value of likelihood function evaluated at the estimates.

For each estimated multivariate volatility model in Table 2, we compute the

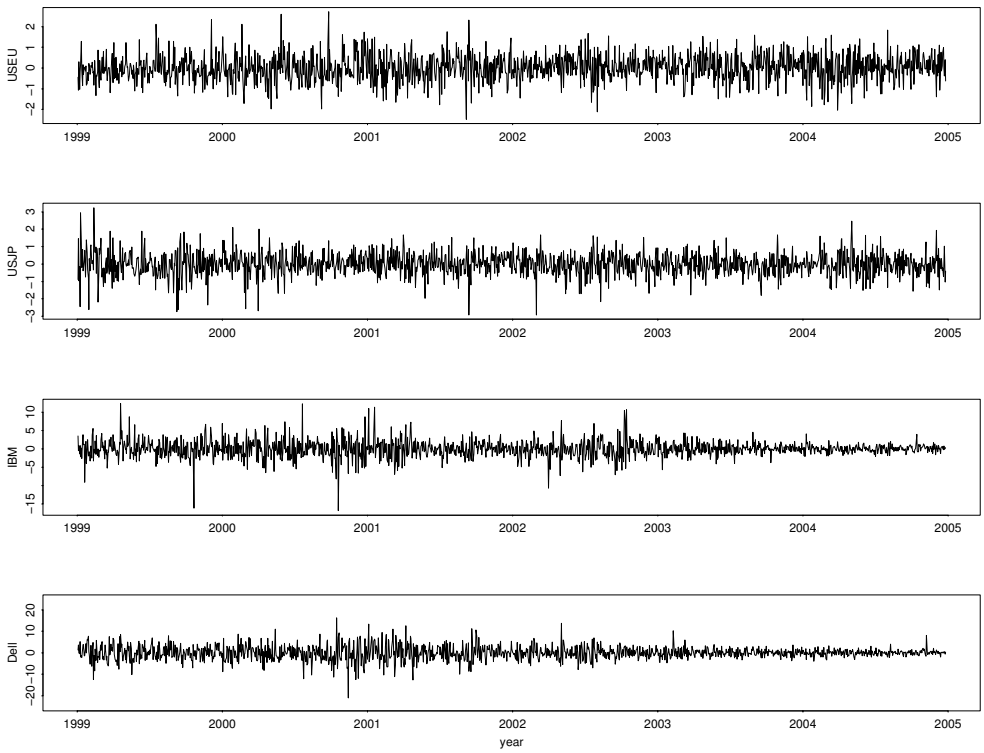


FIG 1. *Time plots of daily return series from January 1999 to December 2004:* (a) *Dollar-Euro exchange rate,* (b) *Dollar-Yen exchange rate,* (c) *IBM stock, and* (d) *Dell stock.*

TABLE 1
Descriptive statistics of daily returns of Example 1. The returns are in percentages, and the sample period is from January 1999 to December 2004 for 1496 *observations*

Asset	USEU	JPUS	IBM	DELL
Mean	0.0091	−0.0059	0.0066	0.0028
Standard error	0.6469	0.6626	5.4280	10.1954
Skewness	0.0342	−0.1674	−0.0530	−0.0383
Excess kurtosis	2.7090	2.0332	6.2164	3.3054
Box-Ljung $Q(12)$	12.5	6.4	24.1	24.1

TABLE 2
Estimation results of multivariate volatility models for Example 1 where $L_{\max}$ *denotes the value of likelihood function evaluated at the estimates,* v *is the degrees of freedom of the multivariate Student-t distribution, and the numbers in parentheses are asymptotic standard errors*

Λ_0	Λ_1	Λ_2	$(v, \theta_1, \theta_2)'$
(a) Full model estimation with $L_{\max} = -9175.80$			
0.0041(0.0033)	0.9701(0.0114)	0.0214(0.0075)	7.8729(0.4693)
0.0088(0.0038)	0.9515(0.0126)	0.0281(0.0084)	0.9808(0.0029)
0.0071(0.0053)	0.9636(0.0092)	0.0326(0.0087)	0.0137(0.0025)
0.0150(0.0136)	0.9531(0.0155)	0.0461(0.0164)	
(b) Restricted model with $L_{\max} = -9176.62$			
0.0066(0.0028)	0.9606(0.0068)	0.0255(0.0068)	7.8772(0.7144)
0.0066(0.0023)		0.0240(0.0059)	0.9809(0.0042)
0.0080(0.0052)		0.0355(0.0068)	0.0137(0.0025)
0.0108(0.0086)		0.0385(0.0073)	
(c) Final restricted model with $L_{\max} = -9177.44$			
0.0067(0.0021)	0.9603(0.0063)	0.0248(0.0048)	7.9180(0.6952)
0.0067(0.0021)		0.0248(0.0048)	0.9809(0.0042)
0.0061(0.0044)		0.0372(0.0061)	0.0137(0.0028)
0.0148(0.0084)		0.0372(0.0061)	
(d) Model with leverage effects, $L_{\max} = -9169.04$			
0.0064(0.0027)	0.9600(0.0065)	0.0254(0.0063)	8.4527(0.7556)
0.0066(0.0023)		0.0236(0.0054)	0.9810(0.0044)
0.0128(0.0055)		0.0241(0.0056)	0.0132(0.0027)
0.0210(0.0099)		0.0286(0.0062)	

standardized residuals as

$$\hat{\epsilon}_t = \widehat{\Sigma}_t^{-1/2} e_t,$$

where $\widehat{\Sigma}_t^{1/2}$ is the symmetric square-root matrix of the estimated volatility matrix $\widehat{\Sigma}_t$. We apply the multivariate Ljung-Box statistics to the standardized residuals $\hat{\epsilon}_t$ and its squared process $\hat{\epsilon}_t^2$ of a fitted model to check model adequacy. For the full model in Table 2(a), we have $Q(10) = 167.79(0.32)$ and $Q(10) = 110.19(1.00)$ for $\hat{\epsilon}_t$ and $\hat{\epsilon}_t^2$, respectively, where the number in parentheses denotes p-value. Clearly, the model adequately describes the first two moments of the return series. For the model in Table 2(b), we have $Q(10) = 168.59(0.31)$ and $Q(10) = 109.93(1.00)$. For the final restricted model in Table 2(c), we obtain $Q(10) = 168.50(0.31)$ and $Q(10) = 111.75(1.00)$. Again, the restricted models are capable of describing the mean and volatility of the return series.

From Table 2, we make the following observations. First, using the likelihood ratio test, we cannot reject the final restricted model compared with the full model. This results in a very parsimonious model consisting of only 9 parameters for the time-varying correlations of the four-dimensional return series. Second, for the two stock return series, the constant terms in Λ_0 are not significantly different from

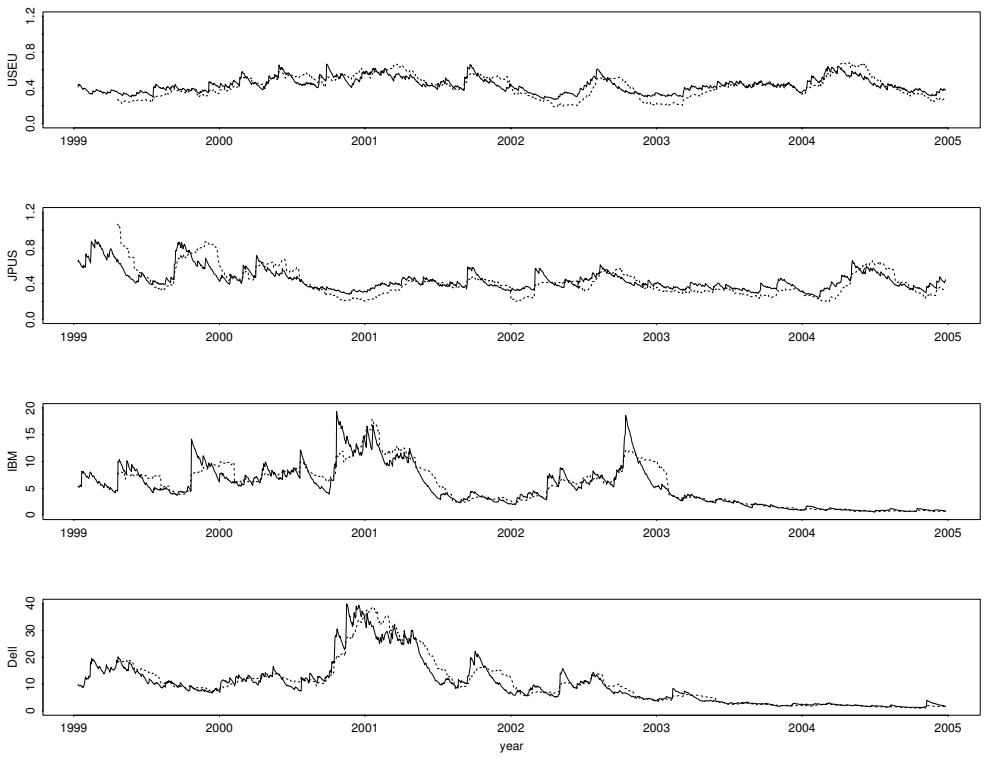


FIG 2. *Time plots of estimated volatility series of four asset returns. The solid line is from the proposed model and the dashed line is from a rolling estimation with window size* 69*:* (a) *Dollar-Euro exchange rate,* (b) *Dollar-Yen exchange rate,* (c) *IBM stock, and* (d) *Dell stock.*

zero and the sum of GARCH parameters is $0.0372 + 0.9603 = 0.9975$, which is very close to unity. Consequently, the volatility series of the two equity returns exhibit IGARCH behavior. On the other hand, the volatility series of the two exchange rate returns appear to have a non-zero constant term and high persistence in GARCH parameters. Third, to better understand the efficacy of the proposed model, we compare the results of the final restricted model with those of rolling estimates. The rolling estimates of covariance matrix are obtained using a moving window of size 69, which is the approximate number of trading days in a quarter. Figure 2 shows the time plot of estimated volatility. The solid line is the volatility obtained by the proposed model and the dashed line is for volatility of the rolling estimation. The overall pattern seems similar, but, as expected, the rolling estimates respond slower than the proposed model to large innovations. This is shown by the faster rise and decay of the volatility obtained by the proposed model. Figure 3 shows the time-varying correlations of the four asset returns. The solid line denotes correlations obtained by the final restricted model of Table 2 whereas the dashed line is for rolling estimation. The correlations of the proposed model seem to be smoother.

Table 2(d) gives the results of a fitted integrated GARCH-type with leverage effects. The leverage effects are statistically significant for equity returns only and are in the form of an IGARCH model. Specifically, the Λ_3 matrix of the correlation equation in Eq. (11) is

$$\Lambda_3 = \text{diag}\{0, 0, (1 - 0.96 - 0.0241), (1 - 0.96 - 0.0286)\} = \text{diag}\{0, 0, 0.0159, 0.0114\}.$$

Although the magnitudes of the leverage parameters are small, but they are statistically significant. This is shown by the likelihood ratio test. Specifically, compared

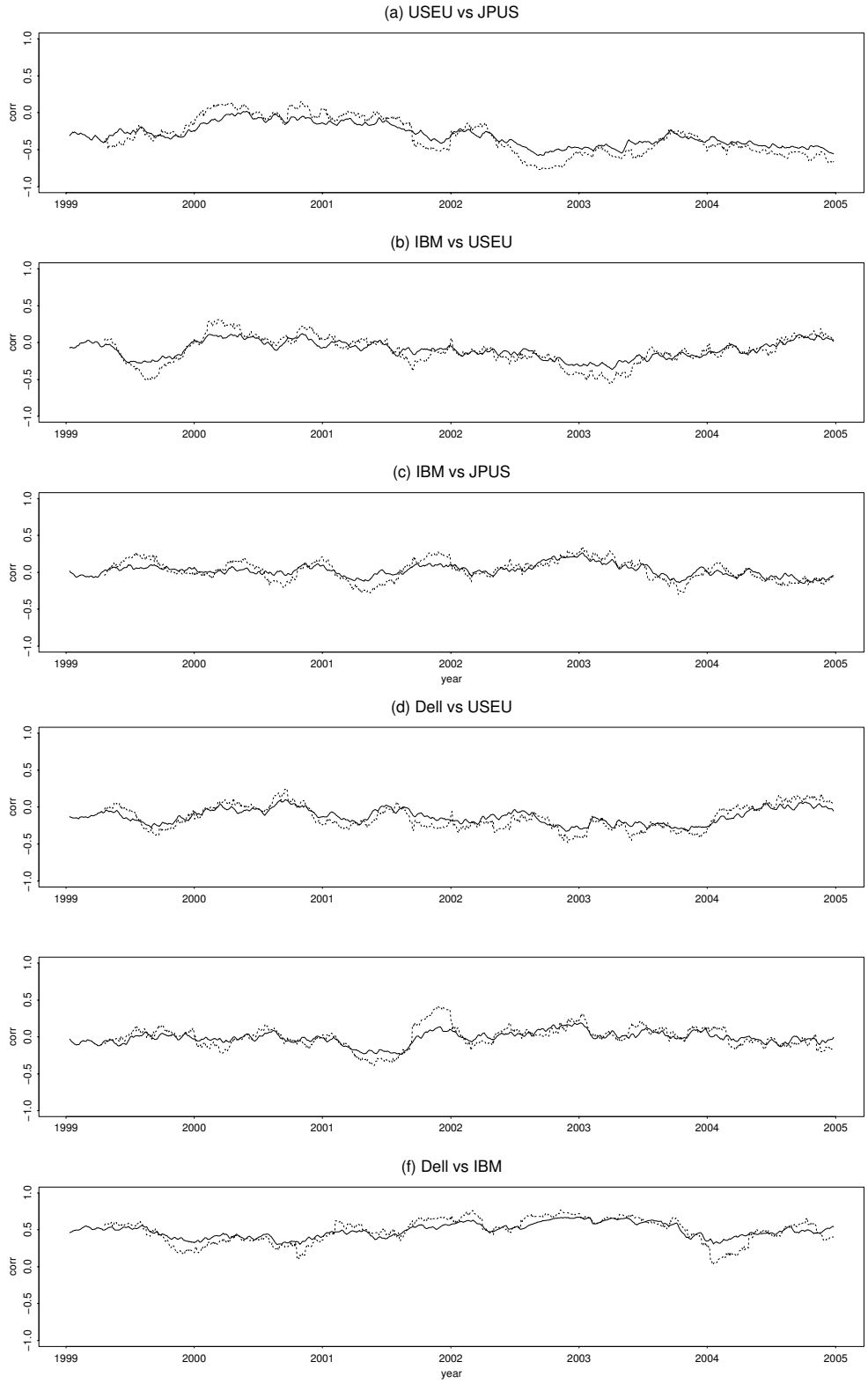


FIG 3. *Time plots of time-varying correlations between the percentage simple returns of four assets from January 1999 to December 2004. The solid line is from the proposed model whereas the dashed line is from a rolling estimation with window size* 69.

the fitted models in Table 2(b) and (d), the likelihood ratio statistic is 15.16, which has a p-value 0.0005 based on the chi-squared distribution with 2 degrees of freedom.

Example 2. In this example, we consider daily simple returns, in percentages, of the S&P 500 index and nine individual stocks from January 1990 to December 2004 for 3784 observations. Thus, we have a 10-dimensional return series. The ten assets are given in Table 3 along with some descriptive statistics. All asset returns have positive excess kurtosis, indicating heavy tails. Except for the stock of General Electrics, return minimums exceed the maximums in modulus, suggesting asymmetry in price changes due to good and bad news.

Sincere there are some minor serial and cross correlations in the 10-dimensional returns, we fit a vector autoregressive model of order 3, i.e. VAR(3), to the data to remove the dynamic dependence and employ the resulting residual series in volatility modeling. See Eq. (8).

We have applied the proposed volatility model in Eqs. (9)- (10) to the residual series of the VAR(3) model. But our subsequently analysis shows that the model with leverage effects in Eq. (11) is preferred based on the likelihood ratio test. Therefore, for simplicity in presentation, we shall only report the results with leverage effects.

Employing the volatility model in Eq. (11) with the correlations in Eq. (10), we found that for the returns of IBM, DELL, GE, and GM stocks the leverage effects follow integrated GARCH models. Consequently, for these four stock returns the leverage parameters are given by

$$\Lambda_{ii,3} = 1 - \Lambda_{ii,1} - \Lambda_{ii,2},$$

where $\Lambda_{ii,j}$ is the ith diagonal element of the matrix Λ_j, $j = 1, 2, 3$. Table 4 shows the parameter estimates of the 10-dimensional volatility model.

For model checking, we use a bootstrap method to generate the critical values of multivariate Ljung-Box statistics for the standardized residuals and their squared series. Specifically, we generate 10,000 realizations each with 3781 observations from the standardized residuals of the fitted model. The bootstrap samples are drawn with replacement. For each realization, we compute the Ljung-Box statistics Q(5), Q(10), and Q(15) of the series and its squared series. Table 5 gives some critical values of the Ljung-Box statistics. For the fitted model, we have $Q(10) = 836.12$ and $Q(15) = 1368.71$ for the standardized residuals and $Q(10) = 1424.64$ and $Q(15) = 1923.75$ for the squared series of standardized residuals. Compared with the critical values in Table 5, the Ljung-Box statistics are not significant at the 1% level. Thus, the fitted model is adequate in modeling the volatility of the 10-dimensional return series. We also applied the AIC criterion to the squared series of standardized residuals. The criterion selected a VAR(0) model, confirming that the fitted multivariate volatility model is adequate.

From the fitted model, we make the following observations. First, except for two marginal cases, all estimates of leverage parameters are statistically significant at the 5% level based on their asymptotic standard errors. The two marginally significant leverage parameters are for BA amd PFE stocks and their t-ratios are 1.65 and 1.92, respectively. Thus, as expected, the leverage effects are positive and most of them are significant. Second, all parameters of the volatility equation are significant. Thus, the model does not contain unnecessary parameters. Third, the model contains 30 parameters. This is very parsimonious for a 10-dimensional return series. Fourth, the correlations evolve slowly with high persistence parameter 0.9864.

TABLE 3
Descriptive statistics of asset returns used in Example 2. Except for the S&P index, tick symbol is used to denote the company. Returns are in percentages

Asset	Mean	St.Error	Skewness	Ex.Kurt.	Minimum	Maximum
S&P	0.038	1.03	−0.018	3.58	−6.87	5.73
IBM	0.066	2.03	0.294	6.32	−15.54	13.16
INTC	0.122	2.82	−0.122	4.17	−22.02	20.12
DELL	0.236	3.49	−0.012	3.45	−25.44	20.76
GE	0.074	1.70	0.176	3.80	−10.67	12.46
BA	0.052	1.98	−0.282	6.08	−17.62	11.63
GM	0.039	2.01	0.111	1.98	−13.53	10.34
JNJ	0.076	1.59	−0.139	4.32	−15.85	8.21
MRK	0.051	1.80	−0.861	14.91	−26.78	9.60
PFE	0.084	1.91	−0.068	1.94	−11.15	9.71

TABLE 4
Parameter estimates of the proposed volatility model with leverage effects for the 10 *asset returns of Example 2. For leverage effects, those estimates without standard errors denote IGARCH constraints*

Λ_1					λI					
Estimate					0.9658					
Std.Err					0.0024					
Λ_2				Diagonal matrix with elements						
Estimate	.0154	.0174	.0168	.0298	.0191	.0206	.0187	.0110	.0128	.0192
Std.Err	.0031	.0026	.0038	.0030	.0029	.0041	.0037	.0038	.0028	.0037
Λ_0				Diagonal matrix with elements						
Estimate	.0077	.0211	.0763	.0170	.0185	.0279	.0342	.0281	.0369	.0309
Std.Err	.0010	.0042	.0121	.0067	.0031	.0054	.0074	.0048	.0061	.0058
Λ_3				Diagonal matrix with elements						
Estimate	.0178	.0168	.0126	.0044	.0152	.0107	.0155	.0210	.0143	.0115
Std.Err	.0049		.0059			.0065		.0064	.0059	.0060
Parameter	v	θ_2	θ_1							
Estimate	9.54	.9864	.0070							
Std.Err	.417	.0016	.0006							

TABLE 5
Critical values of Ljung-Box statistics for 10*-dimensional standardized residual series. The values are obtained by a bootstrap method with* 10,000 *iterations. The sample size of the series is* 3781

	Standardized residuals			Squared standardized residuals		
Statistics	1%	5%	10%	1%	5%	10%
$Q(5)$	576.92	553.68	541.33	915.89	696.82	617.74
$Q(10)$	1109.05	1075.25	1057.94	1150.31	1281.03	1170.12
$Q(15)$	1633.31	1591.61	1571.17	2125.65	1837.28	1713.50

Fifth, the estimate of the degrees of freedom for multivariate Student-t innovation is 9.54, confirming that the returns have heavy tails.

Remark. In this paper, we use a MATLAB program to estimate the proposed multivariate volatility models. The negative log likelihood function is minimized with inequality parameter constraints using the function **fmincon**. Limited experience shows that the results are not sensitive to the initial values, but initial values that are far away from the final estimates do require many more iterations. The

estimation, however, can become difficult if some parameters are approching the boundary of the parameter space. For instance, if there is no leverage effect, then the hessian matrix can be unstable when the leverage parameter is included in the estimation.

5. Extensions and some alternative approaches

In this paper, we consider a simple approach to model multivariate volatilities of asset returns. Unlike other methods available in the literature, the proposed approach estimates the conditional variances and correlations jointly and the resulting volatility matrices are positive definite. The proposed model can handle the leverage effects and is parsimonious. We demonstrated the efficacy of the proposed model by analyzing a 4-dimensional and a 10-dimensional asset return series. The results are encouraging. We also used a bootstrap method to obtain finite-sample critical values for the multivariate Ljung-Box statistics for testing serial and cross correlations of a vector series.

There are possible extensions of the proposed model. For example, Eq. (10) requires that all correlations have the same persistence parameter θ_2. This restriction can be relaxed by letting θ_1 and θ_2 be diagonal matrices of positive real numbers. The model would become

$$R_t = (I - \theta_1^2 - \theta_2^2)\bar{R} + \theta_1 \psi_{t-1} \theta_1 + \theta_2 R_{t-1} \theta_2.$$

Under this model, the ith asset return contributes $\theta_{ii,2}$ to the persistence of correlations. In addition, one can have equality constraints among diagonal elements of each θ_i matrix to keep the model parsimonious.

Some alternative approaches have been considered in the literature to overcome the curse of dimensionality in multivariate volatility modeling. Palandri [7] uses a sequential Cholesky decomposition to build a multivariate volatility of 69 stock returns. The independent component models have also been used to simplify the modeling procedure, e.g., see [6].

References

[1] Bauwens, L., Laurent, S. and Rombouts, J. V. K. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics* **21** 79–109. MR1700749

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

[3] Bollerslev, T., Engle, R. R. and Wooldridge, J. M. (1988). A capital asset pricing model with time varying covariances. *Journal of Political Economy* **96** 116–131.

[4] Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate GARCH models. *Journal of Business and Economic Statistics* **20** 339–350.

[5] Engle, R. F. and Kroner, F. K. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory* **11** 122–150.

[6] Fan, J., Wang, M. and Yao, Q. (2005). Modeling multivariate volatilities via conditionally uncorrelated components. Working Paper, Princeton University.

[7] Palandri, A. (2005). Sequential conditional correlations: Inference and evaluation. Working paper, Department of Economics, Duke University.

[8] Tse, Y. K. and Tsui, A. K. C. (2002). A multivariate GARCH model with time-varying correlations. *Journal of Business and Economic Statistics* **20** 351–362.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 223–235

DOI: 10.1214/074921706000001067

Multi-armed bandit problem with precedence relations

Hock Peng Chan[1,*], Cheng-Der Fuh[2,†] and Inchi Hu[3,‡]

National University of Singapore, National Central University, Academia Sinica and Hong Kong University of Science and Technology

Abstract: Consider a multi-phase project management problem where the decision maker needs to deal with two issues: (a) how to allocate resources to projects within each phase, and (b) when to enter the next phase, so that the total expected reward is as large as possible. We formulate the problem as a multi-armed bandit problem with precedence relations. In Chan, Fuh and Hu (2005), a class of asymptotically optimal arm-pulling strategies is constructed to minimize the shortfall from perfect information payoff. Here we further explore optimality properties of the proposed strategies. First, we show that the efficiency benchmark, which is given by the regret lower bound, reduces to those in Lai and Robbins (1985), Hu and Wei (1989), and Fuh and Hu (2000). This implies that the proposed strategy is also optimal under the settings of aforementioned papers. Secondly, we establish the super-efficiency of proposed strategies when the bad set is empty. Thirdly, we show that they are still optimal with constant switching cost between arms. In addition, we prove that the Wald's equation holds for Markov chains under Harris recurrent condition, which is an important tool in studying the efficiency of the proposed strategies.

1. Introduction

Suppose there are $\mathcal{U} = J_1 + \cdots + J_I$ statistical populations, $\Pi_{11}, \Pi_{12}, \ldots, \Pi_{IJ_I}$. Pulling arm ij once corresponds to taking an observation from population Π_{ij}. The observations from Π_{ij} form a Markov chain on a state space D with transition probability density function $p_{ij}(x, y, \theta)$ with respect to a σ-finite measure Q, where θ is an unknown parameter belonging to a parameter space Θ. The stationary probability distribution for the Markov chain exists and has probability density function $\pi_{ij}(\cdot, \theta)$.

At each step, we are required to sample one of the statistical populations obeying the partial order $ij \preceq i'j' \Leftrightarrow i \leq i'$. An adaptive policy is a sampling rule that dictates, at each step, which population should be sampled based on observations before that step. We can represent a policy as a sequence of random variables $\phi = \{\phi_t | \phi_{t-1} \preceq \phi_t, t = 1, 2, \ldots\}$ taking values in $\{ij | i = 1, \ldots, I; j = 1, \ldots, J_i\}$ such that the event $\{\phi_t = ij\}$ 'take an observation from Π_{ij} at step t' belongs to the σ-field generated by $\phi_1, X_1, \ldots, \phi_{t-1}, X_{t-1}$, where X_t denotes the state of the population being sampled at t-th step.

[1]National University of Singapore, e-mail: stachp@nus.edu.sg
[2]Academia Sinica, e-mail: stcheng@stat.sinica.edu.tw
[3]Hong Kong University of Science and Technology, e-mail: imichu@ust.hk
[*]Research supported by grants from the National University of Singapore.
[†]Research partially supported by the National Science Council of ROC.
[‡]Research partially supported by Hong Kong Research Grant Council.
AMS 2000 subject classifications: primary 62L05; secondary 62N99.
Keywords and phrases: Markov chains, multi-armed bandits, Kullback–Leibler number, likelihood ratio, optimal stopping, scheduling, single-machine job sequencing, Wald's equation.

Let the initial state of Π_{ij} be distributed according to $\nu_{ij}(\cdot;\theta)$. Throughout this paper, we shall use the notation E_θ (P_θ) to denote expectation (probability) with respect to the initial distribution $\nu_{ij}(\cdot;\theta)$; similarly, $E_{\pi(\theta)}$ to denote expectation with respect to the stationary distribution $\pi_{ij}(\cdot;\theta)$. We shall assume that $\mathcal{V}_{ij} = \{x \in D : \nu_{ij}(x;\theta) > 0\}$ does not depend on θ and $v_{ij} := \inf_{x\in\mathcal{V}_{ij}} \inf_{\theta,\theta'\in\Theta}[\nu_{ij}(x;\theta)/\nu_{ij}(x;\theta')] > 0$ for all i, j. Suppose that $\int_{x\in D} |g(x)|\pi_{ij}(x;\theta)Q(dx) < \infty$. Let

$$\mu_{ij}(\theta) = \int_{x\in D} g(x)\pi_{ij}(x;\theta)Q(dx)$$

be the mean reward under stationary distribution π_{ij} when Π_{ij} is sampled once. Let N be the total sample size from all populations, and

$$T_N(ij) = \sum_{t=1}^{N} \mathbf{1}_{\{\phi_t=ij\}} \tag{1.1}$$

be the sample size from Π_{ij} and $\mathbf{1}$ denotes the indicator function. It follows that the total reward equals

$$W_N(\theta) := \sum_{t=1}^{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} E_\theta\{E_\theta[X_t\mathbf{1}_{\{\phi_t=ij\}}|\mathcal{F}_{t-1}]\}. \tag{1.2}$$

In the case of independent rewards, that is, when $p_{ij}(x, y, ;\theta) = p_{ij}(y;\theta)$ for all i, j, x, y and θ, $W_N(\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \mu_{ij}(\theta)E_\theta T_N(ij)$. We shall show in the Appendix that for Markovian rewards, under regularity conditions A3-A4 (see Section 2.1), there exists a constant $C_0 < \infty$ independent of $\theta \in \Theta$, $N > 0$ and the strategy ϕ such that

$$\Big|W_N(\theta) - \sum_{i=1}^{I}\sum_{j=1}^{J_i} \mu_{ij}(\theta)E_\theta T_N(ij)\Big| \le C_0. \tag{1.3}$$

In light of (1.3), maximizing $W_N(\theta)$ is asymptotically equivalent [up to a O(1) term] to minimizing the regret

$$R_N(\theta) := N\mu^*(\theta) - W_N(\theta) = \sum_{ij:\mu_{ij}(\theta)<\mu^*(\theta)} [\mu^*(\theta) - \mu_{ij}(\theta)]E_\theta T_N(ij), \tag{1.4}$$

where $\mu^*(\theta) := \max_{1\le i\le I} \max_{1\le j\le J_i} \mu_{ij}(\theta)$.

Because adaptive strategies ϕ that are optimal for all $\theta \in \Theta$ and large N in general do not exist, we consider the class of all (asymptotically) *uniformly good* adaptive strategies under the partial order constraint $\preceq$, satisfying

$$R_N(\theta) = o(N^\alpha), \qquad \text{for all } \alpha > 0 \text{ and } \theta \in \Theta. \tag{1.5}$$

Such strategies have regret that does not increase too rapidly for any $\theta \in \Theta$. We would like to find a strategy that minimizes the increasing rate of the regret within the class of uniformly good adaptive strategies under the partial order constraint $\preceq$.

The rest of the article is organized as follows. In Section 2, we present the assumptions and introduce the concept of bad sets. The regret lower bound is investigated in Section 3. We also prove that the regret lower bound specializes to other lower bounds obtained by previous authors under less general settings. Section 4 contains the super efficiency result when the bad sets are empty. The optimality of the proposed strategies under constant switching cost is investigated in Section 5. The last section includes the proof of Wald's equation for Markov random walks under Harris recurrence condition.

2. The assumption and bad sets

Denote the Kullback-Leibler information number by

$$I_{ij}(\theta,\theta') = \int_{x\in D}\int_{y\in D} \log\Big[\frac{p_{ij}(x,y;\theta)}{p_{ij}(x,y;\theta')}\Big] p_{ij}(x,y;\theta)\pi_{ij}(x;\theta)Q(dy)Q(dx). \tag{2.1}$$

Then, $0 \le I_{ij}(\theta,\theta') \le \infty$. We shall assume that $I_{ij}(\theta,\theta') < \infty$ for all i,j and $\theta,\theta' \in \Theta$. Let $\mu_i(\theta) = \max_{1\le j\le J_i} \mu_{ij}(\theta)$ be the largest reward in the i-th group of arms, and

$$\Theta_i = \{\theta\in\Theta : \mu_i(\theta) > \mu_{i'}(\theta) \text{ for all } i' < i \text{ and } \mu_i(\theta) \ge \mu_{i'}(\theta) \text{ for all } i' \ge i\} \tag{2.2}$$

be the set of parameter values such that the first optimal job is in group i. Let

$$\Theta_{ij} = \{\theta\in\Theta_i : \mu_{ij}(\theta) = \mu_i(\theta)\} \tag{2.3}$$

be the parameter set such that arm ij is one of the first optimal ones. Each $\theta\in\Theta$ belongs to exactly one Θ_i but may belong to more than one Θ_{ij}. Let

$$\Theta_i^* = \{\theta\in\Theta : \mu_i(\theta) > \mu_{i'}(\theta) \text{ for all } i' \ne i\} \tag{2.4}$$

be the parameter set in which all the optimal arms lie in group i. Clearly, $\Theta_i^* \subset \Theta_i$ but the reverse relation is not necessarily true.

2.1. The assumptions

We now state a set of assumptions that will be used to prove the optimality results. Let Θ be a compact subset of $\mathbf{R}^d$ for some $d \ge 1$.

A1. $\mu_{ij}(\cdot)$ are finite and continuous on Θ for all i,j. Moreover, no arm group is redundant in the sense that $\Theta_i^* \ne \emptyset$ for all $i = 1,\dots,I$.

A2. $\sum_{j=1}^{J_1} I_{1j}(\theta,\theta') > 0$ for all $\theta' \ne \theta$ and $\inf_{\theta'\in\Theta_{ij}} I_{ij}(\theta,\theta') > 0$ for all $1 \le i < I, 1 \le j \le J_i$ and $\theta \in \cup_{\ell>i}\Theta_\ell$.

A3. For each $j = 1,\dots,J_i, i = 1,\dots,I$ and $\theta\in\Theta$, $\{X_{ijt}, t \ge 0\}$ is a Markov chain on a state space D with σ-algebra $\mathcal{D}$, irreducible with respect to a maximal irreducible measure on $(D,\mathcal{D})$ and aperiodic. Furthermore, X_{ijt} is *Harris recurrent* in the sense that there exists a set $G_{ij} \in \mathcal{D}$, $\alpha_{ij} > 0$ and probability measure φ_{ij} on G_{ij} such that $P_{ij}^\theta\{X_{ijt} \in G_{ij} \text{ i.o.} | X_{ij0} = x\} = 1$ for all $x\in D$ and

$$P_{ij}^\theta\{X_{ij1}\in A | X_{ij0} = x\} \ge \alpha_{ij}\varphi_{ij}(A) \quad \text{for all } x\in G_{ij} \text{ and } A\in\mathcal{D}. \tag{2.5}$$

A4. There exist constants $0 < \bar{b} < 1$, $b > 0$ and drift functions $V_{ij} : D \to [1,\infty)$ such that for all $j = 1,\dots,J_i$ and $i = 1,\dots,I$,

$$\sup_{x\in D} |g(x)|/V_{ij}(x) < \infty, \tag{2.6}$$

and for all $x\in D$, $\theta\in\Theta$,

$$P_{ij}^\theta V_{ij}(x) \le (1-\bar{b})V_{ij}(x) + b\mathbf{1}_{G_{ij}}(x), \tag{2.7}$$

where G_{ij} satisfies (2.5) and $P_{ij}^{\theta}V_{ij}(x) = \int_D V_{ij}(y)P_{ij}^{\theta}(x,dy)$. Moreover, we require that

$$\int_D V_{ij}(x)\nu_{ij}(dx;\theta)Q(dx) < \infty \quad \text{and } V_{ij}^* := \sup_{x\in G_{ij}} V_{ij}(x) < \infty. \tag{2.8}$$

Let $\ell_{ij}(x,y;\theta,\theta') = \log[p_{ij}(x,y;\theta)/p_{ij}(x,y;\theta')]$ be the log likelihood ratio between P_{ij}^{θ} and $P_{ij}^{\theta'}$ and $N_\delta(\theta) = \{\theta' : \|\theta - \theta'\| < \delta\}$ a ball of radius δ around θ, where $\|\cdot\|$ denotes Euclidean norm.

A5. There exists $\delta > 0$ such that for all $\theta, \theta' \in \Theta$,

$$K_{\theta,\theta'} := \sup_{x\in D} \frac{E_\theta[\sup_{\tilde\theta\in N_\delta(\theta')} \ell_{ij}^2(X_{ij0}, X_{ij1};\theta,\widetilde{\theta})|X_{ij0} = x]}{V_{ij}(x)} < \infty \tag{2.9}$$

for all $j = 1,\dots,J_i$, $i = 1,\dots,I$. Moreover,

$$\sup_{\tilde\theta\in N_{\delta'}(\theta')} |\ell_{ij}(x,y;\theta',\widetilde{\theta})| \to 0 \text{ as } \delta' \to 0 \tag{2.10}$$

for all $x, y \in D$ and $\theta' \in \Theta$.

Assumption A1 is a mild regularity condition to exclude unrealistic models. A2 is a positive information criterion: the first inequality makes sure that information is available in the first arm group to estimate θ; while the second inequality allows us to collect information in the i-th arm group for moving to the next group when $\theta \in \Theta_\ell$ for some $\ell > i$. Assumption A3 is a recurrence condition and A4 is a drift condition. These two conditions are used to guarantee the stability of the Markov chain so that the strong law of large numbers and Wald's equation hold. A5 is a finite second moment condition that allows us to bound the probability that the MLE of θ lies outside a small neighborhood of θ. This bound is important for us to determine the level of unequal allocation of observations that can be permitted in the testing stage of our procedure. The proof of the asymptotic lower bound in Theorem 1 requires only A1-A3; while additional A4 and A5 are required for the construction of efficient strategies attaining the lower bound.

2.2. Bad sets

The bad set is a useful concept for understanding the learning required within the group containing optimal arms. It is associated with the asymptotic lower bound described in Section 3 and is used explicitly in constructing the asymptotically efficient strategy. For $\theta \in \Theta_\ell$, define $J(\theta) = \{j : \mu^*(\theta) = \mu_{\ell j}(\theta)\}$ as the set of optimal jobs in group ℓ. Hence $\theta \in \Theta_{\ell j}$ if and only if $j \in J(\theta)$. We also define the bad set, the set of 'bad' parameter values associated with θ, as all $\theta' \in \Theta_\ell$ which cannot be distinguished from θ by processing any of the optimal jobs ℓj. Specifically,

$$B_\ell(\theta) = \Big\{\theta' \in \Theta_\ell \setminus \big(\bigcup_{j\in J(\theta)} \Theta_{\ell j}\big) :\ I_{\ell j}(\theta,\theta') = 0 \text{ for all } j \in J(\theta)\Big\}. \tag{2.11}$$

The bad set $B_\ell(\theta)$ is the intersection of two parameter sets. One set consists of parameter values that have different optimal arms from those for θ. The other set contains parameter values that cannot be distinguished from sampling the optimal

arm for θ. When a parameter value is in the intersection, sampling from arms that are non-optimal for θ is required.

We note that if $I_{\ell j}(\theta, \theta') = 0$, then the transition probabilities of $X_{\ell jt}$ are identical under both θ and θ'. If $\theta' \in B_\ell(\theta)$, then by definition, $\theta' \notin \cup_{j \in J(\theta)} \Theta_{\ell j}$ and hence $J(\theta') \cap J(\theta) = \emptyset$. Let $j \in J(\theta)$ and $j' \in J(\theta')$. Then $\mu_{\ell j'}(\theta') > \mu_{\ell j}(\theta') = \mu_{\ell j}(\theta) > \mu_{\ell j'}(\theta)$. Thus

$$I_{\ell j'}(\theta, \theta') > 0 \text{ for all } \theta' \in B_\ell(\theta) \text{ and } j' \in J(\theta'). \tag{2.12}$$

The interpretation of (2.12) is as follows. Although we cannot distinguish θ from $\theta' \in B_\ell(\theta)$ when sampling the optimal arm for θ, we can distinguish them by sampling the optimal job for θ'. This fact explains the necessity of processing non-optimal arms to collect information.

3. The regret lower bound

The following theorem gives an asymptotic lower bound for the regret (1.4) of uniformly good adaptive strategies under the partial order constraint $\preceq$. The proof can be found in [1]. We will discuss the relation of the lower bound with those in [6, 7] and [3].

Theorem 1. *Assume* A1-A3 *and let $\theta \in \Theta_\ell$. For any uniformly good adaptive strategy ϕ under the partial order constraint $\preceq$,*

$$\liminf_{N \to \infty} R_N(\theta)/\log N \geq z(\theta, \ell), \tag{3.1}$$

where $z(\theta, \ell)$ is the minimum value of the following minimization problem.

$$\begin{aligned} &\textit{Minimize} \sum_{i<\ell} \sum_{j=1}^{J_i} [\mu^*(\theta) - \mu_{ij}(\theta)] z_{ij}(\theta) + \sum_{j \notin J(\theta)} [\mu^*(\theta) - \mu_{\ell j}(\theta)] z_{\ell j}(\theta), \\ &\textit{subject to } z_{ij}(\theta) \geq 0, \quad j = 1, \ldots, J_i, \ \textit{if } i < \ell, \ j \notin J(\theta), \ \textit{if } i = \ell, \end{aligned} \tag{3.2}$$

and

$$\begin{cases} \inf_{\theta' \in \Theta_1} \{\sum_{j=1}^{J_1} I_{1j}(\theta, \theta') z_{1j}(\theta)\} \geq 1, \\ \inf_{\theta' \in \Theta_2} \{\sum_{j=1}^{J_1} I_{1j}(\theta, \theta') z_{1j}(\theta) + \sum_{j=1}^{J_2} I_{2j}(\theta, \theta') z_{2j}(\theta)\} \geq 1, \\ \vdots \\ \inf_{\theta' \in \Theta_{\ell-1}} \{\sum_{j=1}^{J_1} I_{1j}(\theta, \theta') z_{1j}(\theta) + \cdots + \sum_{j=1}^{J_{\ell-1}} I_{(\ell-1)j}(\theta, \theta') z_{(\ell-1)j}(\theta)\} \geq 1, \\ \inf_{\theta' \in B_\ell(\theta)} \{\sum_{i<\ell} \sum_{j=1}^{J_i} I_{ij}(\theta, \theta') z_{ij}(\theta) + \sum_{j \notin J(\theta)} I_{\ell j}(\theta, \theta') z_{\ell j}(\theta)\} \geq 1. \end{cases} \tag{3.3}$$

Corollary 1. *When there is only one group of arms, (3.1) reduces to the lower bound (1.11) of Lai and Robbins* [7].

Proof. When there is only group of arms, only the last inequality of (3.3) is needed and it takes the form

$$\inf_{\theta' \in B(\theta)} \sum_{j \notin J(\theta)} I_j(\theta, \theta') z_j(\theta) \geq 1. \tag{3.4}$$

In [7], it is proved that

$$E_\theta T_N(j) \geq \frac{\log N}{I(\theta_j, \theta^*)} \quad \text{for all } j \notin J(\theta), \tag{3.5}$$

where $\theta^* = \max_{1\leq i\leq k} \theta_i$. Note that in [7], all jobs belong to the same family of probability distributions with different parameter values, and thus the KL information number does not depend on the job label but only the parameter value. Let $E_\theta T_N(j)/\log N = z_j(\theta)$, then (3.5) is the same as

$$z_j(\theta) I(\theta_j, \theta^*) \geq 1 \quad \text{for all } j \notin J(\theta). \tag{3.6}$$

We first show that (3.4) $\Rightarrow$ (3.5). Because (3.4) implies that for all $\theta' \in B(\theta)$

$$\sum_{j\notin J(\theta)} I(\theta_j, \theta'_j) z_j(\theta) \geq 1. \tag{3.7}$$

If $\theta' = (\theta'_1, \ldots, \theta'_k) \in B(\theta)$, then $\theta^* = \theta_{j^*} = \theta'_{j^*}$ and $\max_{1\leq i\leq k} \theta'_i > \theta^*$. Suppose we choose a sequence of $\theta' \in B(\theta)$ such that there is only one component θ'_j approaching θ^* from above and other components $\theta'_{j'}$, $j' \notin J(\theta)$, all have the same values as the corresponding components of θ. Taking infimum over this sequence of $\theta' \in B(\theta)$ in (3.7), we obtain (3.6). This complete the proof of (3.4) $\Rightarrow$ (3.5).

To prove (3.5) $\Rightarrow$ (3.4), we assume that (3.4) does not hold. That is, there exist a $\theta' \in B(\theta)$ such that

$$\sum_{j\notin J(\theta)} I(\theta_j, \theta'_j) z_j(\theta) < 1.$$

Because $\theta' \in B(\theta)$, there exists at least one component θ'_{j^*} of θ' such that $\theta'_{j^*} > \theta^*$. Then the preceding inequality and the property of exponential families imply that

$$z_{j^*}(\theta) I(\theta_{j^*}, \theta^*) < z_{j^*}(\theta) I(\theta_{j^*}, \theta'_{j^*}) < 1,$$

and thus (3.6) does not hold. This establishes (3.5) $\Rightarrow$ (3.4) and the proof is complete. □

Corollary 2. *When there is only one arm in each group, then (3.1) reduces to the lower bound (1.17) of Hu and Wei* [6].

Proof. In Hu and Wei [6], the set Θ_i are intervals of $\Re$. Thus the infimum over Θ_i is achieved at the end points of the intervals. Furthermore, because there is only one arm in each group, the bad sets are all empty and therefore the last inequality in (3.3) is not needed. In view of these facts, it is straightforward to show that the systems of inequalities (3.3) reduces to (1.14) of Hu and Wei [6]. The proof is complete. □

Corollary 3. *When there is only one arm in each group, the lower bound (3.1) reduces to (3.2) of Fuh and Hu* [3].

Proof. The assumptions A3 and A4 of Fuh and Hu [3] correspond to the regularity condition A1 and the positive information criterion A2 in Section 2, respectively. The A1, A2 and A5 of Fuh and Hu are essentially the same as Harris recurrence condition A3, the drift condition A4, and the finite second moment condition A5 of this paper, respectively.

Note that the definition of bad sets in [3] is different from that of this paper. In [3], the bad set consists of all those parameter values having optimal arm *not*

in the same group and cannot be distinguished when sampling from the optimal arm. Here the bad set consists of parameter values that has different optimal arm (*but still in the same group*), and cannot be distinguished when sampling from the optimal arm(s). If we adopt the definition (2.11), then it is clear that the bad sets are all empty under the setting of [3].

The infimums in Problem A of Fuh and Hu [3] is taken over the union of Θ_i and the corresponding bad set. Because the bad sets in [3] are all empty as we point out earlier, the infimums is actually taken over Θ_i. With this understanding, it is straightforward to verify that the lower bound (3.1) reduces to (3.2) of Fuh and Hu [3]. □

4. Super efficiency

The strategy in the allocation of the observations is as follows. For the rationale of the proposed strategy and more detailed discussion, please see [1]. Let n_0 and n_1 be positive integers that increase to infinity with respect to N and satisfies $n_0 = o(\log N)$ and $n_1 = o(n_0)$.

1. *Estimation.* Select n_0 observations from each arm in group 1 and let $\widehat{\theta}$ be the maximum likelihood estimate (MLE) of θ defined by

$$L(\theta) = \sum_{j=1}^{J} \sum_{t=1}^{n_0} \log p_{1j}(X_{1j(t-1)}, X_{1jt}; \theta), \quad \widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta). \tag{4.1}$$

Let $\ell = \min\{i : N_{\delta/2}(\widehat{\theta}) \cap \Theta_i \neq \emptyset\}$. Select an adjusted MLE estimate $\widehat{\theta}_a \in N_{\delta/2}(\widehat{\theta}) \cap \Theta_\ell$, (where $\delta \to 0$ as $N \to \infty$ at a rate to be specified in Theorem 1 below), in the following manner. Let $|\cdot|$ denote the number of elements in a finite set and

$$\mathbf{J} = \max\{|J(\theta')| : \theta' \in N_{\delta/2}(\widehat{\theta}) \cap \Theta_\ell\}. \tag{4.2}$$

We require that

$$\widehat{\theta}_a \in H := \{\theta \in N_{\delta/2}(\widehat{\theta}) \cap \Theta_\ell : |J(\theta)| = \mathbf{J}\}. \tag{4.3}$$

The motivation behind considering an adjusted MLE is to estimate $J(\theta)$ and the set Θ_i that θ belongs to consistently. This has implications in the experimentation phase. We note that if $|J(\theta)| > 1$, then $J(\widehat{\theta})$ need not be consistent for $J(\theta)$ and if Θ_i lies on $\Theta_i \setminus \Theta_i^*$ [see (2.2) and (2.4)], then $\widehat{\theta}$ need not be consistently inside Θ_i. Conversely, the probability that $J(\widehat{\theta}_a) = J(\theta)$ and $\widehat{\theta}_a$ lying inside Θ_i tends to 1 as $N \to \infty$.

Let

$$B_\ell(\theta; \delta) = \cup_{\theta' \in H} B_\ell(\theta') \tag{4.4}$$

and let $\{\widehat{z}_{ij}\}_{1 \le i \le \ell, 1 \le j \le J_i}$ minimize

$$\sum_{i<\ell} \sum_{j=1}^{J_i} [\mu^*(\widehat{\theta}_a) - \mu_{ij}(\widehat{\theta}_a)] z_{ij} + \sum_{j \notin J(\widehat{\theta}_a)} [\mu^*(\widehat{\theta}_a) - \mu_{\ell j}(\widehat{\theta}_a)] z_{\ell j} \tag{4.5}$$

subject to the constraints

$$
\begin{cases}
\inf_{\theta' \in \Theta_1} \{\sum_{j=1}^{J_1} I_{1j}(\widehat{\theta}_a, \theta') z_{1j}\} \geq 1, \\
\inf_{\theta' \in \Theta_2} \{\sum_{j=1}^{J_1} I_{1j}(\widehat{\theta}_a, \theta') z_{1j} + \sum_{j=1}^{J_2} I_{2j}(\widehat{\theta}_a, \theta') z_{2j}\} \geq 1, \\
\vdots \\
\inf_{\theta' \in \Theta_{\ell-1}} \{\sum_{j=1}^{J_1} I_{1j}(\widehat{\theta}_a, \theta') z_{1j} + \cdots + \sum_{j=1}^{J_{\ell-1}} I_{(\ell-1)j}(\widehat{\theta}_a, \theta') z_{(\ell-1)j}\} \geq 1, \\
\inf_{\theta' \in B_\ell(\widehat{\theta}; \delta)} \{\sum_{i<\ell} \sum_{j=1}^{J_i} I_{ij}(\widehat{\theta}_a, \theta') z_{ij} + \sum_{j \notin J(\widehat{\theta}_a)} I_{\ell j}(\widehat{\theta}_a, \theta') z_{\ell j}\} \geq 1.
\end{cases} \tag{4.6}
$$

Let $k = 1$.

2. *Experimentation.* If $k \leq \ell$, select $\lfloor \widehat{z}_{kj} \log N \rfloor$ observations from arm kj, where $\lfloor \cdot \rfloor$ denotes the greatest integer function. If $k > \ell$, we skip the experimentation stage. We note that if $B_\ell(\widehat{\theta}; \delta)$ is empty, then the last inequality in (4.6) is automatically satisfied and hence we can select $\widehat{z}_{\ell 1} = \cdots = \widehat{z}_{\ell J_\ell} = 0$. In other words, if $B_\ell(\widehat{\theta}; \delta)$ is empty, then the experimentation stage is also skipped over for $k = \ell$.

3. *Testing.* Start with a full set $\{k1, \ldots, kJ_k\}$ of unrejected jobs. The rejection of a job is based on the following test statistic. Let F_k, $1 \leq k \leq I$, be a probability distribution with positive probability on all open subsets of $\cup_{i=k}^{I} \Theta_i$. Define

$$
U_k(\mathbf{n}; \lambda) = \frac{\int_{\cup_{i=k}^{I} \Theta_i} \prod_{i=1}^{k} \prod_{j=1}^{J_i} \nu_{ij}(X_{ij0}; \theta) \prod_{t=1}^{n_{ij}} p_{ij}(X_{ij(t-1)}, X_{ijt}; \theta) \, dF_k(\theta)}{\prod_{i=1}^{k} \prod_{j=1}^{J_i} \nu_{ij}(X_{ij0}; \lambda) \prod_{t=1}^{n_{ij}} p_{ij}(X_{ij(t-1)}, X_{ijt}; \lambda)} \tag{4.7}
$$

for all $\lambda \in \Theta_k$.

(a) If $\widehat{\theta} \in \cup_{i>k} \Theta_i$: Add one observation from each unrejected job. Reject parameter λ if $U_k(\mathbf{n}; \lambda) \geq N$. Reject a job kj if all $\lambda \in \Theta_{kj}$ have been rejected at some point in the testing stage. If there is a job in group k left unrejected and the total number of observations is less than N, repeat 3(a). Otherwise go to step 4.

(b) If $\widehat{\theta} \in \Theta_k$: Add n_1 observations from each unrejected job kj, $j \in J(\widehat{\theta})$ and one observation from each unrejected job kj, $j \notin J(\widehat{\theta})$. Reject a job kj if all $\lambda \in \Theta_{kj}$ have been rejected at some point in the testing phase. If there is a job in group k left unrejected and the total number of observations is less than N, repeat 3(b). Otherwise, go to step 4.

(c) If $\widehat{\theta} \in \cup_{i<k} \Theta_i$: Adopt the procedure of 3(a).

4. *Moving to the next group and termination.* The strategy terminates once N observations have been collected. Otherwise, if $k < I$, increment k by 1 and go to step 2; if $k = I$, select all remaining observations from a job Ij satisfying $\mu_{Ij}(\widehat{\theta}) = \max_{1 \leq h \leq J_I} \mu_{Ih}(\widehat{\theta})$.

In [1] Theorem 2, it was established that when $B_\ell(\theta)$ is non-empty, then the asymptotic lower bound of the regret is attained with the procedure above. We shall show that the same procedure is not only asymptotically optimal but also the regret from the optimal group will be $o(\log N)$ when $B_\ell(\theta) = \emptyset$ as oppose to $O(\log N)$ when $B_\ell(\theta) \neq \emptyset$. An important key step required in our proof is the consistency result

$$
P_\theta\{B_\ell(\widehat{\theta}, \delta) = \emptyset\} \to 1 \text{ as } N \to \infty \tag{4.8}
$$

under the empty bad set assumption.

Theorem 2. *Let* $\theta \in \Theta_\ell$. *Assume* A1-A5 *and (1.5) . Let* $n_0 \to \infty$ *with* $n_0 = o(\log N)$ *and* $n_1 \to \infty$ *such that* $n_1 = o(n_0)$. *There exists* $\delta(= \delta_N) \downarrow 0$ *as* $N \to \infty$ *such that*

$$P_\theta\{\widehat{\theta} \in \Theta \setminus N_\delta(\theta)\} = o(n_1^{-1}) \text{ as } n_1 \to \infty. \tag{4.9}$$

Moreover, if $B_\ell(\theta) = \emptyset$, *then* (4.8) *holds and*

$$\sum_{j=1}^{J_\ell} E_\theta T_N(\ell j) = o(\log N). \tag{4.10}$$

Hence

$$\lim_{N\to\infty} R_N(\theta)/\log N = z(\theta, \ell). \tag{4.11}$$

Proof. The consistency of $\widehat{\theta}$ in (4.9) follows from A2 and (4.5) of Chan, Fuh and Hu [1]. We shall now prove (4.8). Since $\delta \downarrow 0$ and $\widehat{\theta}$ is consistent for θ, it suffices from the definition of $B_\ell(\theta; \delta)$ in (4.4) to show that there exists $\delta_0 > 0$ such that

$$B_\ell(\widetilde{\theta}) = \emptyset \text{ for all } \widetilde{\theta} \in N_{\delta_0}(\theta) \cap \Theta_\ell \text{ with } |J(\widetilde{\theta})| = \mathbf{J}. \tag{4.12}$$

We observe from the continuity of $\mu_{\ell j}$ that there exists $\delta_1 > 0$ such that $J(\widetilde{\theta}) \subset J(\theta)$ for all $\widetilde{\theta} \in N_{\delta_1}(\theta) \cap \Theta_\ell$. Hence it follows that if $|J(\widetilde{\theta})| = |J(\theta)|$, then it must be true that $J(\widetilde{\theta}) = J(\theta)$. We see from the definition of bad sets in (2.11) that for each $\theta' \in \Theta_\ell \setminus (\cup_{j\in J(\theta)}\Theta_{\ell j})$, $I_{\ell j}(\theta, \theta') > 0$ for some $j \in J(\theta)$ and hence by the continuity of the Kullback-Leibler information, there exists $\delta_2 > 0$ such that $I_{\ell j}(\widetilde{\theta}, \theta') > 0$ whenever $\widetilde{\theta} \in N_{\delta_2}(\theta)$. Select $\delta_0 = \min\{\delta_1, \delta_2\}$. Then (4.12) holds.

We shall next show (4.10). By (4.8) and since the experimentation stage is skipped over when $k = \ell$ and $B_\ell(\widehat{\theta}, \delta) = \emptyset$, it suffices to show that the expected total number of observations taken from inferior arms in the testing stage is $o(\log N)$. Define $p_N = P_\theta\{J(\widehat{\theta}_a) = J(\theta)\}$. Then by (4.3), (4.9) and as $J(\widetilde{\theta}) \subset J(\theta)$ for all $\widetilde{\theta} \in N_{\delta_1}(\theta)$ for some $\delta_1 > 0$, $1 - p_N = o(n_1^{-1})$. By (2.16) and the assumption $B_\ell(\theta) = \emptyset$, at least one optimal arm will provide positive information against each $\theta' \notin \cup_{j\in J(\theta)}\Theta_j$. By A3-A5 and (6.4), (6.5) of Chan, Fuh and Hu [1], (an expected) $O(\log N)$ number of observations from arms with positive information is required to reject each $\theta' \in \Theta_\ell \setminus (\cup_{j\in J(\theta)}\Theta_{\ell j})$. Hence $O(n_1^{-1}\log N)$ number of recursions is involved when $J(\theta) = J(\widehat{\theta}_a)$ because at least n_1 observations in each recursion has positive information. Similarly, $O(\log N)$ recursions is needed when $J(\theta) \neq J(\widehat{\theta}_a)$ because at least one observation in each recursion has positive information. The number of observations from inferior arms in each recursion is $O(1)$ if $J(\widehat{\theta}_a) = J(\theta)$ and $O(n_1)$ otherwise. Hence the expected number of observations from inferior arms during the recursion steps in the testing phase is

$$p_N O(n_1^{-1}\log N) + (1 - p_N)O(n_1 \log N) = o(\log N). \tag{4.13}$$

The asymptotic result (4.11) follows from (4.10) and the proof of Chan, Fuh and Hu [1] Theorem 2. □

For the special case $\ell = 1$, it follows from (4.11) that $R_N(\theta) = o(\log N)$ occurs. In [2] and [10], a uniformly good procedure was proposed that satisfies $R_N(\theta) = O(1)$ when Θ is finite and $I = 1$.

5. The switching cost

Let $a(\theta) > 0$ be the switching cost between two arms and are not both optimal when the underlying parameter is θ. It is assumed here that there is no switching cost when both arms are optimal. Then

$$L_N(\theta) := a(\theta)E_\theta\Big(\sum_{t=1}^{N-1} \mathbf{1}_{\{\phi_t \neq \phi_{t+1}, \min[\mu_{\phi_t}(\theta), \mu_{\phi_{t+1}}(\theta)] < \mu^*(\theta)\}}\Big)$$

is the average switching cost of a procedure. It is also desirable that this cost is asymptotically negligible compared to the regret as $N \to \infty$.

Theorem 3. *Under Assumptions A1 - A5, the strategy ϕ^* has average switching cost*

$$L_N(\theta) = o(\log N) \text{ as } N \to \infty. \tag{5.1}$$

Hence, the strategy is asymptotically optimal when there is switching cost.

Proof. In the estimation stage it is require to take n_0 observations from each arm in group 1. We can take the n_0 observations in batches and switch only $J_1 - 1$ times. Therefore the switching cost from estimation stage is $a(\theta)(J_1 - 1)$. In the experimentation stage, we need to allocate at most $\widehat{z}_{kj} \log N$ observations to arm kj. Again this can be done in batches and thus the switching cost from experimentation stage is at most $a(\theta)(J_k - 1)$. In the testing stage, it is shown in (6.12) of Chan, Fuh and Hu [1], that the expected total number of observations is $o(\log N)$ and thus the switching cost is no more than $o(\log N)$. Adding the switching costs from the estimation, experimentation, and testing stages together, shows that the total cost due to switching is $o(\log N)$. However, the regret lower bound is $O(\log N)$, which implies that the switching cost constitutes a negligible part of the total regret as $n \to \infty$. This completes the proof that the proposed strategy is still asymptotically optimal with constant cost per switch. □

6. Extension of Wald's equation to Markovian rewards

As we will be focusing on a single arm ij and fixed parameters θ_0, θ_q such that $\mu := I_{ij}(\theta_0, \theta_q) > 0$ we will drop some of the references to i, j, θ_0, θ_q and q in this section. This applies also to the notations in assumptions A3-A5. Moreover, we shall use the notation $E(\cdot)$ as a short form of $E_{\theta_0}(\cdot)$ and $E_x(\cdot)$ as a short form of $E_{\theta_0}(\cdot|X_0 = x)$. Let $S_n = \xi_1 + \cdots + \xi_n$, where $\xi_k = \log[p_{ij}(X_{k-1}, X_k; \theta_0)/p_{ij}(X_{k-1}, X_k; \theta_q)]$ has stationary mean μ under P_{θ_0} and let τ be a stopping-time. We shall show that

$$ES_\tau = \mu(E\tau) - E[\gamma(X_\tau)] + E[\gamma(X_0)] \tag{6.1}$$

for some function γ to be specified in Lemma 1. In Lemma 2, we show that the conditions on V in A4-A5 lead to bounds on $\gamma(x)$ and by applying Lemma 3, we obtain

$$E|\gamma(X_\tau)| + E|\gamma(X_0)| = o(E\tau). \tag{6.2}$$

Substituting (6.2) back into (6.1), Wald's equation

$$ES_\tau = [\mu + o(1)]E\tau \tag{6.3}$$

is established for Markovian rewards. Under uniform recurrence condition, Fuh and Lai [4] established Wald's equation based on perturbation theory for the transition operator. The Wald's equation was proved under the assumption that the solution for the Poisson equation exists in [5] based on Poisson equation for the transition operator. In this section, we apply the idea of regeneration epoch to derive the Wald's equation for Markov random walks.

By (2.5), we can augment the Markov additive process and create a split chain containing an atom, so that increments in S_n between visits to the atom are independent. More specifically, we construct stopping-times $0 < \kappa(1) < \kappa(2) < \cdots$ using an auxiliary randomization procedure such that

$$\begin{aligned} &P\{X_{n+1} \in A, \kappa(i) = n+1 | X_n = x, \kappa(i) > n \geq \kappa(i-1)\} \\ &\quad = \begin{cases} \alpha\varphi(A) & x \in G, \\ 0 & \textit{otherwise}. \end{cases} \end{aligned} \tag{6.4}$$

Then by Lemma 3.1 of Ney and Nummelin [9],

(i) $\{\kappa(i+1) - \kappa(i) : i = 1, 2, \ldots\}$ are i.i.d. random variables.
(ii) the random blocks $\{X_{\kappa(i)}, \ldots, X_{\kappa(i+1)-1}\}$, $i = 1, 2, \ldots$, are independent and
(iii) $P\{X_{\kappa(i)} \in A | \mathcal{F}_{\kappa(i)-1}\} = \varphi(A)$, where $\mathcal{F}_n$=σ-field generated by $\{X_0, \ldots, X_n\}$.

By (ii)-(iii), $E_\varphi(S_\kappa - \kappa\mu) = 0$. Define $\kappa = \kappa(1)$. We shall use the notation "n = atom" to denote $n = \kappa(i)$ for some i.

Lemma 1. *Let $\gamma(x) = E_x(S_\kappa - \kappa\mu)$. Then $Z_n = (S_n - n\mu) + \gamma(X_n)$ is a martingale with respect to $\mathcal{F}_n$. Hence (6.1) holds.*

Proof. We can express

$$Z_n = E(S_{U_n} - U_n\mu | \mathcal{F}_n) \quad \text{where } U_n = \inf\{m > n : m = \text{atom}\}.$$

If $X_n = x_n \notin G$, then by (6.4), $U_n > n+1$. Hence $U_{n+1} = U_n$ and

$$E(Z_{n+1} | \mathcal{F}_n) = Z_n \tag{6.5}$$

because $\mathcal{F}_{n+1} \supset \mathcal{F}_n$. If $X_n = x_n \in G$, then by (6.4) and (ii),

$$E(Z_{n+1} | \mathcal{F}_n) - Z_n = E[(S_{U_{n+1}} - S_{U_n}) + (U_{n+1} - U_n) | \mathcal{F}_n] = \alpha E_\varphi(S_\kappa - \kappa\mu) = 0$$

and hence (6.5) also holds. □

Lemma 2. *Under* A3-A5*,*

$$|\gamma(x)| \leq \beta^{-1}[V(x) + b + (V^* + b)V^*(\alpha^{-1} + 1)](K + 1 + |\mu|),$$

where α satisfies (2.5)*, V^* is defined in* A4 *and K is defined in* (2.9)*.*

Proof. By (2.9),

$$V(x) \geq K^{-1}E_x\xi_1^2 \geq K^{-1}(E_x|\xi_1| - 1). \tag{6.6}$$

Let $0 < \sigma(1) < \sigma(2) < \cdots$ be the hitting times of the set G and let $\sigma = \sigma(1)$. Let

$$m_n(A) = E_x\Big[\sum_{n=1}^{\kappa} V(X_n)\mathbf{1}_{\{X_n \in A\}}\Big] \tag{6.7}$$

for all measurable set $A \subset D$. By (2.7),

$$E_x[V(X_n)\mathbf{1}_{\{\sigma \geq n\}}] \leq (1-\beta)E_x[V(X_{n-1})\mathbf{1}_{\{\sigma \geq n-1\}}],\ n \geq 2$$

and

$$E_x[V(X_1)] \leq V(x) + b.$$

Hence by induction,

$$E_x\Big[\sum_{n=1}^{\sigma} V(X_n)\Big] \leq [V(x)+b]\sum_{n=1}^{\infty}(1-\beta)^{n-1} = [V(x)+b]/\beta. \tag{6.8}$$

By (6.7)-(6.8), and as $V \geq 1$,

$$\begin{aligned} m_n(D) &= E_x\Big\{\sum_{n=1}^{\sigma} V(X_n) + \sum_{k=1}^{\infty}\Big[\sum_{n=\sigma(k)+1}^{\sigma(k+1)} V(X_n)\Big]\mathbf{1}_{\{\kappa > \sigma(k)\}}\Big\} \\ &= E_x\Big\{\sum_{n=1}^{\sigma} V(X_n) + \sum_{k=1}^{\infty} E_{X_{\sigma(k)}}\Big[\sum_{n=1}^{\sigma} V(X_n)\Big]\mathbf{1}_{\{\kappa > \sigma(k)\}}\Big\} \\ &\leq \beta^{-1}[V(x)+b] + E_x\Big\{\sum_{k=1}^{\infty}\beta^{-1}[V(X_{\sigma(k)})+b]\mathbf{1}_{\{\kappa > \sigma(k)\}}\Big\} \\ &\leq \beta^{-1}[V(x)+b] + \beta^{-1}(V^*+b)m_n(G). \end{aligned} \tag{6.9}$$

But by (6.4), $m_n(G) \leq V^*(\alpha^{-1}+1)$. Since $\gamma(x) \leq (K+1+|\mu|)m_n(D)$, Lemma 2 holds. □

Let $W_i = |\gamma(X_{\kappa(i)})| + \cdots + |\gamma(X_{\kappa(i+1)-1})|$, for $i \geq 1$. Then by A3-A5, Lemma 4 and its proof, and (i)-(iii), $W_1, W_2, \ldots$ are i.i.d. with finite mean while by (2.8), $W_0 := |\gamma(X_0)| + \cdots + |\gamma(X_{\kappa(1)-1})|$ also has finite mean.

Lemma 3. *Let $M_n = \max_{1 \leq k \leq n} W_k$. Then for any stopping-time τ,*

$$E(M_\tau) = o(E\tau). \tag{6.10}$$

Proof. Let $\delta > 0$ and let $c(= c_\delta) > 0$ be large enough such that $E[(W_1 - c)^+] \leq \delta$. We shall show that

$$Z_n = (M_n \vee c) - n\delta$$

is a supermartingale. Indeed for any $\lambda \geq 0$,

$$E[M_{n+1} \vee c | M_n \vee c = c + \lambda] = c + \lambda + E[(W_{n+1} - c - \lambda)^+] \leq c + \lambda + \delta$$

and the claim is shown. Hence $EZ_\tau \leq EZ_0 = c$ and it follows that $E(M_\tau) \leq E(M_\tau \vee c) \leq \delta(E\tau) + c$. Lemma 3 then follows by letting $\delta \downarrow 0$. □

7. Appendix

Proof. Proof of (1.3) Let X_{ijt} denotes the tth observation taken from arm ij. Then

$$\Big|W_N(\theta) - \sum_{i=1}^{I}\sum_{j=1}^{J_i}\mu_{ij}(\theta)E_\theta T_N(ij)\Big| \leq \sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{t=1}^{\infty}|E_\theta g(X_{ijt}) - \mu_{ij}(\theta)|. \tag{7.1}$$

For any signed measure λ on $(D, \mathcal{D})$, let

$$\|\lambda\|_{V_{ij}} = \sup_{h:|h|\le V_{ij}} \Big| \int h(x)\lambda(dx) \Big|. \tag{7.2}$$

It follows from Meyn and Tweedie ([8], p.367 and Theorem 16.0.1) that under A3 and the geometric drift condition (2.7),

$$\omega_{ij} := \sup_{\theta\in\Theta, x\in D} \sum_{t=1}^{\infty} \|P_{ijt}^{\theta}(x,\cdot) - \pi_{ij}(\theta)\|_{V_{ij}}/V_{ij}(x) < \infty, \tag{7.3}$$

where $P_{ijt}^{\theta}(x,\cdot)$ denotes the distribution of X_{ijt} conditioned on $X_{ij0} = x$ and $\pi_{ij}(\theta)$ denotes the stationary distribution of X_{ijt} under parameter θ. By (2.6), there exists $\kappa > 0$ such that $\kappa|g(x)| \le V_{ij}(x)$ for all $x \in D$ and hence it follows from (7.2) and (7.3) that

$$\kappa \sum_{t=1}^{\infty} |E_{\theta,x} g(X_{ijt}) - \mu_{ij}(\theta)| \le \omega_{ij} V_{ij}(x), \tag{7.4}$$

where $E_{\theta,x}$ denotes expectation with respect to P_θ and intial distribution $X_{ij0} = x$.

In general, for any initial distribution $\nu_{ij}(\cdot;\theta)$, it follows from (2.8) and (7.4) that

$$\sum_{t=1}^{\infty} |E_\theta g(X_{ijt}) - \mu_{ij}(\theta)| \le \int \sum_{t=1}^{\infty} |E_{\theta,x} g(X_{ijt}) - \mu_{ij}(\theta)| \nu_{ij}(x;\theta) Q(dx) < \infty$$

uniformly over $\theta \in \Theta$ and hence (1.3) follows from (7.1). □

References

[1] Chan, H. P., Fuh, C. D. and Hu, I. (2005). Optimal strategies for a class of sequential control problems with precedence relations. *Annals of Statistics*, to appear.

[2] Feldman, D. (1962). Contributions to the "two-armed bandit" problem. *Annals of Mathematical Statistics* **33** 847–856.

[3] Fuh, C.D. and Hu, I. (2000). Asymptotically efficient strategies for a stochastic scheduling problem with order constraints. *Annals of Statistics* **28** 1670–1695.

[4] Fuh, C. D. and Lai, T. L. (1998). Wald's equations, first passage times and moments of ladder variables in Markov random walks. *Journal of Applied Probability* **35** 566–580.

[5] Fuh, C. D. and Zhang, C. H. (2000). Poisson equation, moment inequalities and r-quick convergence for Markov random walks. *Stochastic Processes and their Applications* **87** 53–67.

[6] Hu, I. and Wei, C. Z. (1989). Irreversible adaptive allocation rules. *Annals of Statistics***17** 801–823.

[7] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6** 4–22.

[8] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chain and Stochastic Stability*. Springer-Verlag, New York.

[9] Ney, P. and Nummelin, E. (1987). Markov additive processes I: eigenvalue properties and limit theorems. *Annals of Probability* **15** 561–592.

[10] Rodman, L. (1978). On the many-armed bandit problem. *Annals of Probability* **6** 491–498.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 236–244

DOI: 10.1214/074921706000001076

Poisson process approximation: From Palm theory to Stein's method

Louis H. Y. Chen[1,*] **and Aihua Xia**[2,†]

National University of Singapore and University of Melbourne

Abstract: This exposition explains the basic ideas of Stein's method for Poisson random variable approximation and Poisson process approximation from the point of view of the immigration-death process and Palm theory. The latter approach also enables us to define local dependence of point processes [Chen and Xia (2004)] and use it to study Poisson process approximation for locally dependent point processes and for dependent superposition of point processes.

1. Poisson approximation

Stein's method for Poisson approximation was developed by Chen [13] which is based on the following observation: a nonnegative integer valued random variable W follows Poisson distribution with mean λ, denoted as $\mathrm{Po}(\lambda)$, if and only if

$$\mathbb{E}\{\lambda f(W+1) - Wf(W)\} = 0$$

for all bounded $f\colon \mathbb{Z}_+ \to \mathbb{R}$, where $\mathbb{Z}_+ := \{0,1,2,\ldots\}$. Heuristically, if $\mathbb{E}\{\lambda f(W+1) - Wf(W)\} \approx 0$ for all bounded $f\colon \mathbb{Z}_+ \to \mathbb{R}$, then $\mathcal{L}(W) \approx \mathrm{Po}(\lambda)$. On the other hand, as our interest is often on the difference $\mathbb{P}(W \in A) - \mathrm{Po}(\lambda)(A) = \mathbb{E}[\mathbf{1}_A(W) - \mathrm{Po}(\lambda)(A)]$, where $A \subset \mathbb{Z}_+$ and $\mathbf{1}_A$ is the indicator function on A, it is natural to relate the function $\lambda f(w+1) - wf(w)$ with $\mathbf{1}_A(w) - \mathrm{Po}(\lambda)(A)$, leading to the Stein equation:

$$\lambda f(w+1) - wf(w) = \mathbf{1}_A(w) - \mathrm{Po}(\lambda)(A). \tag{1}$$

If the equation permits a bounded solution f_A, then

$$\mathbb{P}(W \in A) - \mathrm{Po}(\lambda)(A) = \mathbb{E}\{\lambda f_A(W+1) - Wf_A(W)\};$$

and

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) &:= \sup_{A \subset \mathbb{Z}_+} |\mathbb{P}(W \in A) - \mathrm{Po}(\lambda)(A)| \\ &= \sup_{A \subset \mathbb{Z}_+} |\mathbb{E}\{\lambda f_A(W+1) - Wf_A(W)\}|. \end{aligned}$$

[1]Institute for Mathematical Sciences, National University of Singapore, 3 Prince George's Park, Singapore 118402, Republic of Singapore, e-mail: matchyl@nus.edu.sg

[2]Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia, e-mail: xia@ms.unimelb.edu.au

*Supported by research grant R-155-000-051-112 of the National University of Singapore.

†Supported by the ARC Centre of Excellence for Mathematics and Statistics of Complex Systems.

AMS 2000 subject classifications: primary 60–02; secondary 60F05, 60G55.

Keywords and phrases: immigration-death process, total variation distance, Wasserstein distance, locally dependent point process, hard core process, dependent superposition of point processes, renewal process.

As a special case in applications, we consider independent Bernoulli random variables $X_1, \cdots, X_n$ with $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p_i$, $1 \le i \le n$, and $W = \sum_{i=1}^n X_i$, $\lambda = \mathbb{E}(W) = \sum_{i=1}^n p_i$. Since

$$\mathbb{E}[Wf(W)] = \sum_{i=1}^n \mathbb{E}[X_i f(W)] = \sum_{i=1}^n p_i \mathbb{E} f(W_i + 1),$$

where $W_i = W - X_i$, we have

$$\begin{aligned}\mathbb{E}\{\lambda f_A(W+1) - W f_A(W)\} &= \sum_{i=1}^n p_i \mathbb{E}\left[f_A(W+1) - f_A(W_i+1)\right] \\ &= \sum_{i=1}^n p_i^2 \mathbb{E}\Delta f_A(W_i+1),\end{aligned}$$

where $\Delta f_A(i) = f_A(i+1) - f_A(i)$. Further analysis shows that $|\Delta f_A(w)| \le \frac{1-e^{-\lambda}}{\lambda}$ (see [6] for an analytical proof and [26] for a probabilistic proof). Therefore

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le \left(1 \wedge \frac{1}{\lambda}\right) \sum_{i=1}^n p_i^2.$$

Barbour and Hall [7] proved that the lower bound of $d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda))$ above is of the same order as the upper bound. Thus this simple example of Poisson approximation demonstrates how powerful and effective Stein's method is. Furthermore, it is straightforward to use Stein's method to study the quality of Poisson approximation to the sum of dependent random variables which has many applications (see [18] or [8] for more information).

2. Poisson process approximation

Poisson process plays the central role in modeling the data on occurrence of rare events at random positions in time or space and is a building block for many other models such as Cox processes, marked Poisson processes (see [24]), compound Poisson processes and Lévy processes. To adapt the above idea of Poisson random variable approximation to Poisson process approximation, we need a probabilistic interpretation of Stein's method which was introduced by Barbour [4]. The idea is to split f by defining $f(w) = g(w) - g(w-1)$ and rewrite the Stein equation (1) as

$$\mathcal{A}g(w) \colon = \lambda[g(w+1) - g(w)] + w[g(w-1) - g(w)] = \mathbf{1}_A(w) - \mathrm{Po}(\lambda)(A), \tag{2}$$

where $\mathcal{A}$ is the generator of an immigration-death process $Z_w(t)$ with immigration rate λ, unit per capita death rate, $Z_w(0) = w$, and stationary distribution $\mathrm{Po}(\lambda)$. The solution to the Stein equation (2) is

$$g_A(w) = -\int_0^\infty \mathbb{E}[\mathbf{1}_A(Z_w(t)) - \mathrm{Po}(\lambda)(A)]dt. \tag{3}$$

This probabilistic approach to Stein's method has made it possible to extend Stein's method to higher dimensions and process settings. To this end, let Γ be a compact metric space which is the carrier space of the point processes being approximated. Suppose d_0 is a metric on Γ which is bounded by 1 and ρ_0 is a pseudo-metric on Γ

which is also bounded by 1 but generates a weaker topology. We use δ_x to denote the point mass at x, let $\mathcal{X} := \{\sum_{i=1}^k \delta_{\alpha_i} : \alpha_1, \ldots, \alpha_k \in \Gamma, k \geq 1\}$, $\mathcal{B}(\mathcal{X})$ be the Borel σ–algebra generated by the weak topology ([23], pp. 168–170): a sequence $\{\xi_n\} \subset \mathcal{X}$ converges weakly to $\xi \in \mathcal{X}$ if $\int_\Gamma f(x)\xi_n(dx) \to \int_\Gamma f(x)\xi(dx)$ as $n \to \infty$ for all bounded continuous functions f on Γ. Such topology can also be generated by the metric d_1 defined below (see [27], Proposition 4.2). A *point process* on Γ is defined as a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ (see [23], p. 13). We use Ξ to stand for a point process on Γ with finite intensity measure $\boldsymbol{\lambda}$ which has total mass $\lambda := \boldsymbol{\lambda}(\Gamma)$, where $\boldsymbol{\lambda}(A) = \mathbb{E}\Xi(A)$, for all Borel set $A \subset \Gamma$. Let $\mathrm{Po}(\boldsymbol{\lambda})$ denote the distribution of a Poisson process on Γ with intensity measure $\boldsymbol{\lambda}$.

Since a point process on Γ is an $\mathcal{X}$-valued random element, the key step of extending Stein's method from one dimensional Poisson approximation to higher dimensions and process settings is, instead of considering $\mathbb{Z}_+$-valued immigration-death process, we now need an immigration-death process defined on $\mathcal{X}$. More precisely, by adapting (2), Barbour and Brown [5] define the Stein equation as

$$\begin{aligned} \mathcal{A}g(\xi) &:= \int_\Gamma [g(\xi + \delta_x) - g(\xi)]\boldsymbol{\lambda}(dx) + \int_\Gamma [g(\xi - \delta_x) - g(\xi)]\xi(dx) \\ &= h(\xi) - \mathrm{Po}(\boldsymbol{\lambda})(h), \end{aligned} \tag{4}$$

where $\mathrm{Po}(\boldsymbol{\lambda})(h) = \mathbb{E}h(\zeta)$ with $\zeta \sim \mathrm{Po}(\boldsymbol{\lambda})$. The operator $\mathcal{A}$ is the generator of an $\mathcal{X}$-valued immigration-death process $Z_\xi(t)$ with immigration intensity $\boldsymbol{\lambda}$, unit per capita death rate, $Z_\xi(0) = \xi \in \mathcal{X}$, and stationary distribution $\mathrm{Po}(\boldsymbol{\lambda})$. Its solution is

$$g_h(\xi) = -\int_0^\infty \mathbb{E}[h(Z_\xi(t)) - \mathrm{Po}(\boldsymbol{\lambda})(h)]dt, \tag{5}$$

(see [5]).

To measure the error of approximation, we use Wasserstein pseudo-metric which has the advantage of allowing us to lift the carrier space to a bigger carrier space. Of course, other metrics such as the total variation distance can also be considered and the only difference is to change the set of test functions h. Let

$$\rho_1\left(\sum_{i=1}^m \delta_{x_i}, \sum_{j=1}^n \delta_{y_j}\right) := \begin{cases} 1 & \text{if } m \neq n, \\ \min_\pi \frac{1}{m}\sum_{i=1}^m \rho_0(x_i, y_{\pi(i)}) & \text{if } m = n \geq 1, \\ 0 & \text{if } n = m = 0, \end{cases}$$

where the minimum is taken over all permutations π of $\{1, 2, \ldots, m\}$. Clearly, ρ_1 is a metric (resp. pseudo-metric) if ρ_0 is a metric (resp. pseudo-metric) on $\mathcal{X}$. Set

$$\mathcal{H} = \{h \text{ on } \mathcal{X} : |h(\xi_1) - h(\xi_2)| \leq \rho_1(\xi_1, \xi_2) \text{ for all } \xi_1, \xi_2 \in \mathcal{X}\}.$$

For point processes Ξ_1 and Ξ_2, define

$$\rho_2(\mathcal{L}(\Xi_1), \mathcal{L}(\Xi_2)) := \sup_{h \in \mathcal{H}} |\mathbb{E}h(\Xi_1) - \mathbb{E}h(\Xi_2)|,$$

then ρ_2 is a metric (resp. pseudo-metric) on the distributions of point processes if ρ_1 is a metric (resp. pseudo-metric). In summary, we defined a Wasserstein pseudo-metric on the distributions of point processes on Γ through a pseudo-metric on Γ as shown in the following chart:

Carrier space Γ		Configuration space $\mathcal{X}$		Space of the distributions of point processes
ρ_0 (≤ 1)	$\longrightarrow$	ρ_1 (≤ 1)	$\longrightarrow$	ρ_2 (≤ 1)

As a simple example, we consider a Bernoulli process defined as

$$\Xi = \sum_{i=1}^{n} X_i \delta_{\frac{i}{n}},$$

where, as before, $X_1, \ldots, X_n$ are independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p_i$, $1 \leq i \leq n$. Then Ξ is a point process on carrier space $\Gamma = [0,1]$ with intensity measure $\boldsymbol{\lambda} = \sum_{i=1}^{n} p_i \delta_{\frac{i}{n}}$. With the metric $\rho_0(x,y) = |x-y|\colon = d_0(x,y)$, we denote the induced metric ρ_2 by d_2. Using the Stein equation (4), we have

$$\begin{aligned}
&\mathbb{E}h(\Xi) - \mathrm{Po}(\boldsymbol{\lambda})(h) \\
&\quad = \mathbb{E}\left\{\int_\Gamma [g_h(\Xi + \delta_x) - g_h(\Xi)]\boldsymbol{\lambda}(dx) + \int_\Gamma [g_h(\Xi - \delta_x) - g_h(\Xi)]\Xi(dx)\right\} \\
&\quad = \sum_{i=1}^{n} p_i \mathbb{E}\left\{[g_h(\Xi + \delta_{\frac{i}{n}}) - g_h(\Xi)] - [g_h(\Xi_i + \delta_{\frac{i}{n}}) - g_h(\Xi_i)]\right\} \\
&\quad = \sum_{i=1}^{n} p_i^2 \mathbb{E}\left\{[g_h(\Xi_i + 2\delta_{\frac{i}{n}}) - g_h(\Xi_i + \delta_{\frac{i}{n}})] - [g_h(\Xi_i + \delta_{\frac{i}{n}}) - g_h(\Xi_i)]\right\},
\end{aligned}$$

where $\Xi_i = \Xi - X_i \delta_{\frac{i}{n}}$. It was shown in [27], Proposition 5.21, that

$$\sup_{h \in \mathcal{H}, \alpha, \beta \in \Gamma} |g_h(\xi + \delta_\alpha + \delta_\beta) - g_h(\xi + \delta_\alpha) - g_h(\xi + \delta_\beta) + g_h(\xi)| \leq \frac{3.5}{\lambda} + \frac{2.5}{|\xi| + 1}, \tag{6}$$

where, and in the sequel, $|\xi|$ is the total mass of ξ, $\lambda = \boldsymbol{\lambda}(\Gamma) = \sum_{i=1}^{n} p_i$. Hence

$$\begin{aligned}
d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) &= \sup_{h \in \mathcal{H}} |\mathbb{E}h(\Xi) - \mathrm{Po}(\boldsymbol{\lambda})(h)| \\
&\leq \sum_{i=1}^{n} p_i^2 \left(\frac{3.5}{\lambda} + \mathbb{E}\frac{2.5}{\sum_{1 \leq j \leq n, j \neq i} X_j + 1}\right) \qquad (7) \\
&\leq \frac{6}{\lambda - \max_{1 \leq i \leq n} p_i} \sum_{i=1}^{n} p_i^2
\end{aligned}$$

since

$$\begin{aligned}
\mathbb{E}\frac{1}{\sum_{1 \leq j \leq n, j \neq i} X_j + 1} &= \mathbb{E}\int_0^1 z^{\sum_{1 \leq j \leq n, j \neq i} X_j} dz \\
&= \int_0^1 \prod_{1 \leq j \leq n, j \neq i} [zp_j + (1 - p_j)] dz \\
&\leq \int_0^1 \prod_{1 \leq j \leq n, j \neq i} e^{-p_j(1-z)} dz = \int_0^1 e^{-(\lambda - p_j)(1-z)} dz \leq \frac{1}{\lambda - p_j},
\end{aligned}$$

(see [27], pp. 167–168). Since $d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \geq d_{TV}(\mathcal{L}(|\Xi|), \mathrm{Po}(\lambda))$ and the lower bound of $d_{TV}(\mathcal{L}(|\Xi|), \mathrm{Po}(\lambda))$ is of the same order as $\frac{1}{\lambda}\sum_{i=1}^{n} p_i^2$ [7], the bound in (7) is of the optimal order.

3. From Palm theory to Stein's method

Barbour's probabilistic approach to Stein's method is based on the conversion of a first order difference equation to a second order difference equation. In this section, we take another approach to Stein's method from the point of Palm theory. The connection between Stein's method and Palm theory has been known to many others (e.g., T. C. Brown (personnel communication), [9]) and the exposition here is mainly based on [14] and [27].

There are two properties which distinguish a Poisson process from other processes: independent increments and the number of points on any bounded set follows Poisson distribution. Hence, a Poisson process can be thought as a process pieced together by lots of independent "Poisson components" (if the location is an atom, the "component" will be a Poisson random variable, but if the location is diffuse, then the "component" is either 0 or 1) ([27], p. 121). Consequently, to specify a Poisson process N, it is sufficient to check that "each component" $N(d\alpha)$ is Poisson and independent of the others, that is $\mathbb{E}\{[\mathbb{E}N(d\alpha)]g(N+\delta_\alpha) - N(d\alpha)g(N)\} = 0$, which is equivalent to

$$\frac{\mathbb{E}[g(N)N(d\alpha)]}{\mathbb{E}N(d\alpha)} = \mathbb{E}g(N+\delta_\alpha), \tag{8}$$

for all bounded function g on $\mathcal{X}$ and all $\alpha \in \Gamma$ (see [27], p. 121). To make the heuristic argument rigorous, one needs the tools of Campbell measures and Radon-Nikodym derivatives ([23], p. 83).

In general, for each point process Ξ with finite mean measure $\boldsymbol{\lambda}$, we may define the Campbell measure $C(B, M) = \mathbb{E}[\Xi(B)\mathbf{1}_{\Xi\in M}]$ for all Borel $B \subset \Gamma$, $M \in \mathcal{B}(\mathcal{X})$. This measure is finite and admits the following disintegration:

$$C(B, M) = \int_B Q_s(M)\boldsymbol{\lambda}(ds), \tag{9}$$

or equivalently,

$$Q_s(M) = \frac{\mathbb{E}[\Xi(ds)\mathbf{1}_{\Xi\in M}]}{\boldsymbol{\lambda}(ds)},\ M \in \mathcal{B}(\mathcal{X}),\ s \in \Gamma\ \boldsymbol{\lambda}\ a.s.,$$

where $\{Q_s,\ s \in \Gamma\}$ are probability measures on $\mathcal{B}(\mathcal{X})$ ([23], p. 83 and p. 164) and are called *Palm distributions*. Moreover, (9) is equivalent to that, for any measurable function $f\colon \Gamma \times \mathcal{X} \to \mathbb{R}_+$,

$$\mathbb{E}\left(\int_B f(\alpha, \Xi)\Xi(d\alpha)\right) = \int_B \int_{\mathcal{X}} f(\alpha, \xi)Q_\alpha(d\xi)\boldsymbol{\lambda}(d\alpha) \tag{10}$$

for all Borel set $B \subset \Gamma$. A point process Ξ_α (resp. $\Xi_\alpha - \delta_\alpha$) on Γ is called a *Palm process* (resp. *reduced Palm process*) of Ξ at location α if it has the Palm distribution Q_α and, when Ξ is a simple point process (a point process taking values 0 or 1 at each location), the Palm distribution $\mathcal{L}(\Xi_\alpha)$ can be interpreted as the conditional distribution of Ξ given that there is a point of Ξ at α. It follows from (10) that the Palm process satisfies

$$\mathbb{E}\int_\Gamma f(\alpha, \Xi)\Xi(d\alpha) = \mathbb{E}\int_\Gamma f(\alpha, \Xi_\alpha)\boldsymbol{\lambda}(d\alpha)$$

for all bounded measurable functions f on $\Gamma \times \mathcal{X}$. In particular, Ξ is a Poisson process if and only if

$$\mathcal{L}(\Xi_\alpha) = \mathcal{L}(\Xi + \delta_\alpha),\ \boldsymbol{\lambda}\ a.s.$$

where the extra point δ_α is due to the "Poisson property" of $\Xi\{\alpha\}$, and $\Xi_\alpha|_{\Gamma\setminus\{\alpha\}}$ has the same distribution as $\Xi|_{\Gamma\setminus\{\alpha\}}$ because of independent increments. Here $\xi|_A$ stands for the point measure restricted to $A \subset \Gamma$ ([23], p. 12). In other words, $\Xi \sim \mathrm{Po}(\boldsymbol{\lambda})$ if and only if

$$\mathbb{E}\left\{\int_\Gamma f(\alpha, \Xi + \delta_\alpha)\boldsymbol{\lambda}(d\alpha) - \int_\Gamma f(\alpha, \Xi)\Xi(d\alpha)\right\} = 0,$$

for a sufficiently rich class of functions f, so we define

$$Df(\xi) \colon = \int_\Gamma f(x, \xi + \delta_x)\boldsymbol{\lambda}(dx) - \int_\Gamma f(x, \xi)\xi(dx).$$

If $\mathbb{E}Df(\Xi) \approx 0$ for an appropriate class of test functions f, then $\mathcal{L}(\Xi_\alpha)$ is close to $\mathcal{L}(\Xi + \delta_\alpha)$, which means that $\mathcal{L}(\Xi)$ is close to $\mathrm{Po}(\boldsymbol{\lambda})$ under the metric or pseudo-metric specified by the class of test functions f.

If f_g is a solution of

$$Df(\xi) = g(\xi) - \mathrm{Po}(\boldsymbol{\lambda})(g),$$

then a distance between $\mathcal{L}(\Xi)$ and $\mathrm{Po}(\boldsymbol{\lambda})$ is measured by $|\mathbb{E}Df_g(\Xi)|$ over the class of functions g.

From above analysis, we can see that there are many possible solutions f_g for a given function g. The one which admits an immigration-death process interpretation is by setting

$$f(x, \xi) = h(\xi) - h(\xi - \delta_x),$$

so that Df takes the following form:

$$Df(\xi) = \int_\Gamma [h(\xi + \delta_x) - h(\xi)]\boldsymbol{\lambda}(dx) + \int_\Gamma [h(\xi - \delta_x) - h(\xi)]\xi(dx) = \mathcal{A}h(\xi),$$

where $\mathcal{A}$ is the same as the generator defined in section 2.

4. Locally dependent point processes

We say a point process Ξ is *locally dependent* with neighborhoods $\{A_\alpha \subset \Gamma \colon\ \alpha \in \Gamma\}$ if $\mathcal{L}(\Xi|_{A_\alpha^c}) = \mathcal{L}(\Xi_\alpha|_{A_\alpha^c}),\ \alpha \in \Gamma\ \boldsymbol{\lambda}\ a.s.$

The following theorem is virtually from Corollary 3.6 in [14] combined with the new estimates of Stein's factors in [27], Proposition 5.21.

Theorem 4.1. *If Ξ is a point process on Γ with finite intensity measure $\boldsymbol{\lambda}$ which has the total mass λ and locally dependent with neighborhoods $\{A_\alpha \subset \Gamma \colon\ \alpha \in \Gamma\}$. Then*

$$\rho_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le \mathbb{E}\int_{\alpha\in\Gamma}\left(\frac{3.5}{\lambda} + \frac{2.5}{|\Xi^{(\alpha)}| + 1}\right)(\Xi(A_\alpha) - 1)\Xi(d\alpha)$$
$$+\mathbb{E}\int_{\alpha\in\Gamma}\int_{\beta\in A_\alpha}\left(\frac{3.5}{\lambda} + \frac{2.5}{|\Xi_\beta^{(\alpha)}| + 1}\right)\boldsymbol{\lambda}(d\alpha)\boldsymbol{\lambda}(d\beta),$$

where $\Xi^{(\alpha)} = \Xi|_{A_\alpha^c}$ and $\Xi_\beta^{(\alpha)} = \Xi_\beta|_{A_\alpha^c}$.

Remark. The error bound is a "correct" generalization of $\frac{1}{\lambda}\sum_{i=1}^n p_i^2$ with the Stein factor $\frac{1}{\lambda}$ replaced by a nonuniform bound.

5. Applications

5.1. Matérn hard core process on $\mathbb{R}^d$

A Matérn hard core process Ξ on compact $\Gamma \subset \mathbb{R}^d$ is a model for particles with repulsive interaction. It assumes that points occur according to a Poisson process with uniform intensity measure on Γ. The configurations of Ξ are then obtained by deleting any point which is within distance r of another point, irrespective of whether the latter point has itself already been deleted [see Cox & Isham [17], p. 170].

The point process is locally dependent with neighborhoods $\{B(\alpha, 2r)\colon\ \alpha \in \Gamma\}$, where $B(\alpha, s)$ is the ball centered at α with radius s. Let $\boldsymbol{\lambda}$ be the intensity measure of Ξ, $d_0(\alpha, \beta) = \min\{|\alpha - \beta|, 1\}$, then

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) = O\left(\frac{\mu \mathrm{Vol}(B(0,1))(2r)^d}{\mathrm{Vol}(\Gamma)}\right),$$

where μ is the mean of the total number of points of the original Poisson process (see [14], Theorem 5.1).

5.2. Palindromes in a genome

Let $\{I_i\colon\ 1 \le i \le n\}$ be locally dependent Bernoulli random variables, $\{U_i\colon\ 1 \le i \le n\}$ be independent Γ-valued random elements which are also independent of $\{I_i\colon\ 1 \le i \le n\}$, set $\Xi = \sum_{i=1}^n I_i \delta_{U_i}$, then Ξ is a point process on Γ. For $U_i = i/n$ this point process models palindromes in a genome where I_i represents whether a palindrome occurs at i/n. The point process can also be used to describe the vertices in a random graph.

In general, the U_i's could take the same value and one cannot tell which U_i and therefore which I_i contributes to the value. To overcome this difficulty we lift the process up to a point process $\Xi' = \sum_{i=1}^n I_i \delta_{(i,U_i)}$ on a larger space $\Gamma' = \{1, 2, \ldots, n\} \times \Gamma$. The metric d_0 becomes a pseudo-metric ρ_0, that is, $\rho_0((i,s), (j,t)) = d_0(s,t)$, and Ξ' a locally dependent process (see [14], section 4). It turns out that the Poisson process approximation of $\Xi = \sum_{i=1}^n I_i \delta_{U_i}$ is a special case of the following section.

5.3. Locally dependent superposition of point processes

Since the publication of the Grigelionis Theorem [20] which states that the superposition of independent sparse point processes on carrier space $\mathbb{R}_+$ is close to a Poisson process, there has been a lot of study on the weak convergence of point processes to a Poisson process under various conditions (see, e.g., [16, 19, 21] and [10]). Extensions to dependent superposition[1] of sparse point processes have been carried out in [1–3, 11, 22]. Schuhmacher [25] considered the Wasserstein distance between the weakly dependent superposition of sparse point processes and a Poisson process.

Let Γ be a compact metric space, $\{\Xi_i\colon i \in \mathcal{I}\}$ be a collection of point processes on Γ with intensity measures $\boldsymbol{\lambda}_i$, $i \in \mathcal{I}$. Define $\Xi = \sum_{i \in \mathcal{I}} \Xi_i$ with intensity measure

[1] We use "(resp. locally, weakly) dependent superposition of point processes" to mean that the point processes are (resp. locally, weakly) dependent among themselves.

$\boldsymbol{\lambda} = \sum_{i\in\mathcal{I}} \boldsymbol{\lambda}_i$. Assume $\{\Xi_i\colon\ i \in \mathcal{I}\}$ are locally dependent: that is, for each $i \in \mathcal{I}$, there exists a neighbourhood $A_i \subset \mathcal{I}$ such that $i \in A_i$ and Ξ_i is independent of $\{\Xi_j\colon\ j \notin A_i\}$.

The locally dependent point process $\Xi = \sum_{i=1}^n I_i\delta_{U_i}$ can be regarded as a locally dependent superposition of point processes defined above.

Theorem 5.1 ([15]). *With the above setup, $\lambda = \boldsymbol{\lambda}(\Gamma)$, we have*

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le \mathbb{E}\sum_{i\in\mathcal{I}}\left(\frac{3.5}{\lambda} + \frac{2.5}{|\Xi^{(i)}|+1}\right)\int_\Gamma d_1'(\mathbf{V}_i, \mathbf{V}_{i,\alpha})\boldsymbol{\lambda}_i(d\alpha) + \sum_{i\in\mathcal{I}}\left(\frac{3.5}{\lambda} + \mathbb{E}\frac{2.5}{|\Xi^{(i)}|+1}\right)\mathbb{E}\int_\Gamma d_1'(\Xi_i, \Xi_{i,\alpha})\boldsymbol{\lambda}_i(d\alpha),$$

where $\Xi^{(i)} = \sum_{j\notin A_i}\Xi_j$, $\mathbf{V}_i = \sum_{j\in A_i\setminus\{i\}}\Xi_j$, $\Xi_{i,\alpha}$ is the reduced Palm process of Ξ_i at α,

$$\mathbb{P}(\mathbf{V}_{i,\alpha} \in M) = \frac{\mathbb{E}[\Xi_i(d\alpha)\mathbf{1}_{\mathbf{V}_i\in M}]}{\mathbb{E}\Xi_i(d\alpha)} \text{ for all } M \in \mathcal{B}(\mathcal{X})$$

and

$$d_1'(\xi_1, \xi_2) = \min_{\pi\colon \text{permutations of } \{1,\dots,m\}}\sum_{i=1}^n d_0(y_i, z_{\pi(i)}) + (m-n)$$

for $\xi_1 = \sum_{i=1}^n \delta_{y_i}$ and $\xi_2 = \sum_{i=1}^m \delta_{z_i}$ with $m \ge n$ [Brown & Xia [12]].

Corollary 5.2 ([14]). *For $\Xi = \sum_{i\in\mathcal{I}} I_i\delta_{U_i}$ and $\lambda = \sum_{i\in\mathcal{I}} p_i$ defined in section 5.2,*

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le \mathbb{E}\sum_{i\in\mathcal{I}}\sum_{j\in A_i\setminus\{i\}}\left(\frac{3.5}{\lambda} + \frac{2.5}{V_i+1}\right)I_iI_j + \sum_{i\in\mathcal{I}}\sum_{j\in A_i}\left(\frac{3.5}{\lambda} + \mathbb{E}\left[\frac{2.5}{V_i+1}\middle| I_j = 1\right]\right)p_ip_j,$$

where $V_i = \sum_{j\notin A_i} I_j$.

Corollary 5.3 ([15]). *Suppose that $\{\Xi_i\colon\ 1 \le i \le n\}$ are independent renewal processes on $[0,T]$ with the first arrival time of Ξ_i having distribution G_i and its inter-arrival time having distribution F_i, and let $\Xi = \sum_{i\in\mathcal{I}}\Xi_i$ and $\boldsymbol{\lambda}$ be its intensity measure, then*

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le \frac{6\sum_{i=1}^n [2F_i(T) + G_i(T)]G_i(T)/(1-F_i(T))^2}{\sum_{i=1}^n G_i(T) - \max_j \frac{G_j(T)}{1-F_j(T)}}.$$

References

[1] Banys, R. (1975). The convergence of sums of dependent point processes to Poisson processes. (Russian. Lithuanian, English summary) *Litovsk. Mat. Sb.* **15** 11–23, 223.

[2] Banys, R. (1985). A Poisson limit theorem for rare events of a discrete random field. (Russian. English, Lithuanian summary) *Litovsk. Mat. Sb.* **25** 3–8.

[3] Banys, R. (1980). On superpositions of random measures and point processes. *Mathematical Statistics and Probability Theory.* Proc. Sixth Internat. Conf., Wisla, 1978, *Lecture Notes in Statist.* **2**. Springer, New York-Berlin, pp. 26–37.

[4] Barbour, A. D. (1988). Stein's method and Poisson process convergence. *J. Appl. Probab.* **25** (A) 175–184.
[5] Barbour, A. D. and Brown, T. C. (1992). Stein's method and point process approximation. *Stoch. Procs. Applics* **43** 9–31.
[6] Barbour, A. D. and Eagleson, G. K. (1983). Poisson approximation for some statistics based on exchangeable trials. *Adv. Appl. Prob.* **15** 585–600.
[7] Barbour, A. D. and Hall, P. (1984). On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95** 473–480.
[8] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation.* Oxford Univ. Press.
[9] Barbour, A. D. and Månsson, M. (2002). Compound Poisson process approximation. *Ann. Probab.* **30** 1492–1537.
[10] Brown, T. C. (1978). A martingale approach to the Poisson convergence of simple point processes. *Ann. Probab.* **6** 615–628.
[11] Brown, T. C. (1979). Position dependent and stochastic thinning of point processes. *Stoch. Procs. Applics* **9** 189–193.
[12] Brown, T. C. and Xia, A. (1995). On metrics in point process approximation. *Stochastics and Stochastics Reports* **52** 247–263.
[13] Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545.
[14] Chen, L. H. Y. and Xia, A. (2004). Stein's method, Palm theory and Poisson process approximation. *Ann. Probab.* **32** 2545–2569.
[15] Chen, L. H. Y. and Xia, A. (2006). Poisson process approximation for dependent superposition of point processes. *(preprint).*
[16] Çinlar, E. (1972). Superposition of point processes. *Stochastic Point Processes: Statistical Analysis, Theory, and Applications.* Conf., IBM Res. Center, Yorktown Heights, NY, 1971. Wiley-Interscience, New York, pp. 549–606.
[17] Cox, D. R. and Isham, V. (1980). *Point Processes.* Chapman & Hall.
[18] Erhardsson, T. (2005). Poisson and compound Poisson approximation. In: *An Introduction to Stein's Method*, Eds. A. D. Barbour and L. H. Y. Chen. World Scientific Press, Singapore, pp. 61–113.
[19] Goldman, J. R. (1967). Stochastic point processes: limit theorems. *Ann. Math. Statist.* **38** 771–779.
[20] Grigelionis, B. (1963). On the convergence of sums of random step processes to a Poisson process. *Theor. Probab. Appl.* **8** 177–182.
[21] Jagers, P. (1972). On the weak convergence of superpositions of point processes. *Z. Wahrsch. verw. Geb.* **22** 1–7.
[22] Kallenberg, O. (1975). Limits of compound and thinned point processes. *J. Appl. Probab.* **12** 269–278.
[23] Kallenberg, O. (1983). *Random Measures.* Academic Press, London.
[24] Kingman, J. F. C. (1993). *Poisson processes.* Oxford University Press.
[25] Schuhmacher, D. (2005). Distance estimates for dependent superpositions of point processes. *Stoch. Procs. Applics* **115** 1819–1837.
[26] Xia, A. (1999). A probabilistic proof of Stein's factors. *J. Appl. Probab.* **36** 287–290.
[27] Xia, A. (2005). Stein's method and Poisson process approximation. In: *An Introduction to Stein's Method*, Eds. A. D. Barbour and L. H. Y. Chen. World Scientific Press, Singapore, pp. 115–181.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 245–256

DOI: 10.1214/074921706000001085

Statistical modeling for experiments with sliding levels

Shao-Wei Cheng[1], C. F. J. Wu[2] and Longcheen Huwang[3]

Academia Sinica, Georgia Institute of Technology and National Tsing-Hua University

Abstract: Design of experiment with related factors can be implemented by using the technique of sliding levels. Taguchi (1987) proposed an analysis strategy by re-centering and re-scaling the slid factors. Hamada and Wu (1995) showed via counter examples that in many cases the interactions cannot be completely eliminated by Taguchi's strategy. They proposed an alternative method in which the slid factors are modeled by nested effects. In this work we show the inadequacy of both methods when the objective is response prediction. We propose an analysis method based on a response surface model, and demonstrate its superiority for prediction. We also study the relationships between these three modeling strategies.

1. Introduction

In many investigations, the experimenters can choose an appropriate interval as the experimental range for each factor. The overall experimental region is then the cube formed by the tensor product of these intervals. Such an experimental region is called *regular*. However, when some of the factors are related, an appropriate experimental region becomes irregular and thus cannot be constructed in the usual manner. Factors are called *related* when the desirable experimental region of some factors depends on the level settings of other factors. Design of experiments with related factors can be implemented by using the technique of *sliding levels* proposed by Taguchi [7]. It has been used in practice for a long time but has received scant attention in the statistical literature. Some examples can be found in [2, 6, 7]. Li et al. [4] proposed a two-stage strategy for the sliding-level experiments whose desriable experimental region is unknown and needs to be explored during the experiment. Here the use of sliding is more complicated due to its engineering needs.

In this article we study the situations in which only one factor is chosen to be slid. This article is organized as follows. In Section 2, we will review the existing work on the sliding level technique and show the inadequacy of these methods when the objective of the experiment is response prediction. In Section 3, we will propose an analysis method based on a response surface model, and demonstrate its superiority for prediction. In Section 4, an illustration with a welding experiment will be given. In Section 5, some results are presented based on a comparison between the response surface approach and Taguchi's approach. A summary is given in the last section.

[1]Academia Sinica, Institute of Statistical Science, Taipei, Taiwan, e-mail: swcheng@stat.sinica.edu.tw

[2]Georgia Institute of Technology, School of Industrial and Systems Engineering, Atlanta, GA 30332-0205, USA, e-mail: jeffwu@isye.gatech.edu

[3]National Tsing-Hua University, Institute of Statistics, Hsinchu, Taiwan, e-mail: huwang@stat.nthu.edu.tw

AMS 2000 subject classifications: primary 62K15, 62K20; secondary 62P30.

Keywords and phrases: irregular experimental region, nested effect, re-centering and re-scaling, response prediction, response surface modeling, robust parameter design.

2. Existing approaches

Taguchi [7] justified the use of sliding levels by the rationales of *bad region avoidance* and *interaction elimination*. The analysis strategy in his approach for sliding levels can be interpreted as a *re-centering* and *re-scaling* (RCRS) transformation, which transforms an irregular experimental region into a regular one as shown in Fig 1. In data analysis, this transformation is essentially to code the factor levels by regarding the slid factor as a non-slid factor. For example, $(+1, -1)$ is used for the conditional low and high levels respectively in a two-level slid factors, and $(-1, 0, +1)$ for the conditional low, median, and high levels respectively in a three-level slid factors with equally spaced levels. Consider two factors A and B, in which there are several sliding levels for B at each level of A. It is easy to show that an interaction in the original factor space is eliminated after RCRS only if the relationship between the mean response $E(y)$ and factors A and B satisfies the relationship:

$$E(y) = g_1(x_A) + g_2\left[\frac{x_B - c_B(x_A)}{r_B(x_A)}\right], \tag{1}$$

where g_1 and g_2 are two arbitrary functions and the c's and r's represent the centering and scaling constants with those for factor B depending on factor A. Furthermore, to eliminate the interaction between A and B for mean response satisfying (1), a proper choice of sliding levels based on c's and r's is required. As pointed out via a counter example by Hamada and Wu [3], inadequately locating the sliding levels will not remove the interaction. Similarly, an inadequate choice of scale will not eliminate the interaction neither.

One can infer that the sliding levels must be chosen properly in order to eliminate a potentially removable interaction. To achieve this, one has to know the exact relationship between the factors and the mean response $E(y)$. Because this relationship is not available, an experiment needs to be carried out. Therefore the advantage of interaction elimination by using sliding levels is questionable. Even though the related factors' interactions can be removed by proper centering and scaling, important information like robustness may be masked (see [3], for more details).

Hamada and Wu [3] proposed a *nested-effects modeling* (NEM) approach by using a regression model with nested effects. Because the actual settings of the slid factor are different at each level combination of its related factors, sliding-levels designs

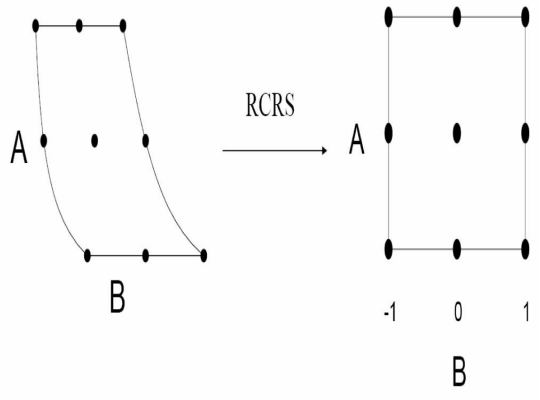


FIG 1. *Re-centering and re-scaling transformation of experimental region.*

can be viewed as nested designs. Hence, one can model the effect of the nested (slid) factor separately at each level combination of its related factors, i.e., the effects of the slid factor are defined conditional on the level combinations of its related factor. Consider the case of two related factors where factor B's levels depend on A's. The factor A can be either qualitative or quantitative. For qualitative A, Hamada and Wu [3] proposed analyzing the effect of B at each level of A. If B is quantitative with more than two levels, the linear and quadratic effects of B at the ith level of A (denoted by $B_l|A_i$ and $B_q|A_i$) should be analyzed. Furthermore, the effects of factor A are analyzed as well. For instance, if A is qualitative with three levels, the two contrasts $A_{1,2}$ and $A_{1,3}$ can be considered, where $A_{i,j}$ represents the contrast between levels i and j of A, i.e., it denotes the difference between the average responses over the conditional levels of B at level i of A and those at level j of A. Because the levels of B vary with the level of A, this is different from the usual meaning of $A_{i,j}$ in factorial designs with regular experimental region, where the same set of levels of B is used for i and j. If A is quantitative, the linear and quadratic effects of A (i.e., A_l and A_q in the linear-quadratic system defined in [8]) should be substituted for $A_{1,2}$ and $A_{1,3}$. The same reasoning will show that the meanings of A_l and A_q are again different from the usual ones.

The analysis using a regression model with nested effects resolves the problem that the sliding-levels design may not eliminate the interaction between related factors. It also provides more insight into the response-factor relationship and directly accounts for the relationship between related factors, which can be used to choose optimum factor levels. However, as far as response prediction is concerned, the nested effects analysis is incapable of accomplishing the task for quantitative A.

When A is a quantitative factor, we may need to predict the response at a setting whose value of A, say x_A^*, is not included in the experimental plan. To achieve this, we need to have a fitted model of B at $A = x_A^*$. However, such a model is not available in the NEM approach because an NEM offers fitted models of B *only* for each levels of A and x_A^* is not one of the levels in the experiment. Therefore, response prediction at x_A^* cannot be achieved in an NEM approach. Because the effects of B are defined and analyzed conditional on A in an NEM, A is treated like a qualitative factor in the analysis about B. This results in the difficulty of response prediction at x_A^*. Turning to the RCRS approach for performing prediction, we have to know the centering and scaling constants of B at x_A^*, i.e., $c_B(x_A^*)$ and $r_B(x_A^*)$, so that B can be appropriately transformed at $A = x_A^*$ before substituting into the fitted RCRS model. However, both $c_B(x_A^*)$ and $r_B(x_A^*)$ may not be available to the investigators. In the next section, we shall propose an analysis method based on the response surface methodology and demonstrates its superiority for response prediction over the two existing approaches.

3. An analysis strategy based on response surface modeling

Response surface modeling (RSM) is an effective tool for building empirical models for the input and output variables in an experiment. In RSM, a true model is often expressed as $y = f(x_1, x_2, \ldots, x_k) + \epsilon$, where y is the observed response, f a function of k quantitative factors $x_1, x_2, \ldots, x_k$, and ϵ an error term. For simplicity, the lowest level of a factor is coded as -1 and the highest level as $+1$. The function f represents the response surface, which depicts the true relationship between the response and factors. Because the form of f is often unknown, RSM replaces and approximates f by a polynomial model of degree d in the x_i's. In practical applications, d is often

chosen to be one or two, and three when the response surface is expected to be more complicated and there are sufficient degrees of freedom. Fourth and higher degree polynomials are rarely used because they are not as effective as semi-parametric or nonparametric models. Further discussion on the response surface methodology can be found in [1] and [5].

In a sliding-level experiment, the adequate experimental region, denoted by R_E, usually has an irregular shape in contrast to the regular region in conventional factorial experiments. In such circumstances, the RSM can still be applied by first finding a cuboidal region that covers exactly the R_E as follows. For each factor, let its lowest actual setting be coded as -1 and the highest actual setting as $+1$. Other settings of the factor is then proportionally coded according to their distances from the lowest one. In this coding, the cuboidal region $[-1, +1]^k$ is the smallest cube to cover the R_E. We call $[-1, +1]^k$ the *modeling region* and denote it by R_M. The RSM can then be applied in the modeling region to develop an empirical model. Unlike factorial designs with regular experimental region, the design points in a sliding-level experiment do not spread uniformly on the whole modeling region. Because there are no design points located in $R_M \backslash R_E$, we have no information about the response surface over $R_M \backslash R_E$. Therefore, the fitted model may fit well only in R_E, but not in the whole R_M. Another issue concerns the choice of appropriate polynomial models for the approximation of the true response surface. For sliding-level experiments, should we still use a dth-order polynomial model? This will be further explained later. When a fitted model is obtained, prediction can be easily done in the RSM approach. Its prediction is an interpolation in R_E but an extrapolation in $R_M \backslash R_E$. An illustration of the RSM strategy will be given in Section 4.

Consider a nine-run experiment with factors A and B, in which A has three levels and conditional on each level of A, B has three sliding levels. The NEM for the experiment can be written as:

$$f(B|A_i) = b_0^i + b_l^i(B_l|A_i) + b_q^i(B_q|A_i), \quad i = -1, 0, 1, \tag{2}$$

where b_0^i, b_l^i and b_q^i are the conditional constant, linear, and quadratic main effects of B given $A = i$. Because A has three levels, the NEM has nine effects and therefore is saturated. On the other hand, a second-order RSM model for the experiment has only six effects. Because the NEM is saturated, it is clear that the RSM model is a submodel of the NEM. In other words, we can impose some constraints on the parameters of the NEM to obtain the RSM model. To find these constraints, we re-parameterize the NEM in (2) in terms of the coding based on the RSM as follows:

$$f(x_B|x_A) = \alpha_{x_A} + \beta_{x_A} x_B + \gamma_{x_A} x_B^2, \quad x_A = -1, 0, 1, \tag{3}$$

where x_B is coded according to the RSM approach but nested on x_A, and α_{x_A}, β_{x_A}, and γ_{x_A} are the zero-order, first-order, and second-order effects of B conditional on $A = x_A$, respectively. Note that for $x_A = i$, x_B is a linear transformation of $B_l|A_i$, and x_B^2 is a linear combination of 1, $B_l|A_i$, and $B_q|A_i$. By equating the NEM in (3) and the following second-order RSM model:

$$f(x_A, x_B) \approx \lambda_0 + \lambda_1 x_A + \lambda_2 x_B + \lambda_{11} x_A^2 + \lambda_{22} x_B^2 + \lambda_{12} x_A x_B, \tag{4}$$

we obtain the following relationships:

$$\begin{aligned} \alpha_{x_A} &= \lambda_0 + \lambda_1 x_A + \lambda_{11} x_A^2, \\ \beta_{x_A} &= \lambda_2 + \lambda_{12} x_A, \\ \gamma_{x_A} &= \lambda_{22}. \end{aligned} \tag{5}$$

The equations in (5) indicate that the three conditional second-order effects of B (i.e., γ_i's) must be identical in the second-order RSM model, which save two degrees of freedom; the three conditional first-order effects of B (i.e., β_{x_A}'s) must satisfy a linear constraint, which save one degree of freedom. The saving of three degrees of freedom explains why the RSM model has three parameters fewer than the NEM.

If the restrictions on β_{x_A}'s and γ_{x_A}'s in (5) are considered to be too rigid, we can add more parameters in the RSM model so that the corresponding β_{x_A}'s and γ_{x_A}'s can be free of the constraints as shown by the following relationships:

$$\begin{aligned} \alpha_{x_A} &= \lambda_0 + \lambda_1 x_A + \lambda_{11} x_A^2, \\ \beta_{x_A} &= \lambda_2 + \lambda_{12} x_A + \lambda_{112} x_A^2, \\ \gamma_{x_A} &= \lambda_{22} + \lambda_{122} x_A + \lambda_{1122} x_A^2. \end{aligned} \tag{6}$$

The resulting RSM model will be:

$$\begin{aligned} f(x_A, x_B) \approx{}& \lambda_0 + \lambda_1 x_A + \lambda_{11} x_A^2 \\ &+ (\lambda_2 + \lambda_{12} x_A + \lambda_{112} x_A^2) x_B \\ &+ (\lambda_{22} + \lambda_{122} x_A + \lambda_{1122} x_A^2) x_B^2 \\ ={}& \lambda_0 + \lambda_1 x_A + \lambda_2 x_B + \lambda_{12} x_A x_B + \lambda_{11} x_A^2 + \lambda_{22} x_B^2 \\ &+ \lambda_{112} x_A^2 x_B + \lambda_{122} x_A x_B^2 + \lambda_{1122} x_A^2 x_B^2. \end{aligned}$$

By adding three higher-order effects $x_A^2 x_B$, $x_A x_B^2$, and $x_A^2 x_B^2$ in the model (4), the RSM model has the same number of parameters and same capacity of estimation as the saturated NEM.

From the previous explanation, it is observed that the conventional RSM approach of using a dth-order model can be inappropriate for data from sliding-level experiments. For example, if a second-order model is adopted, some implicit constraints that can be impractical are placed on β_{x_A}'s and γ_{x_A}'s. However, if the experimenter would like to use a more complicated model, such as a third-order model, there are not enough degrees of freedom for estimating all parameters.

Another interesting observation about the relationship between NEM and RSM model can be obtained from the equations in (6). Consider, for example, the three conditional zero-order effects, α_{x_A}'s. They are individually estimated at each level of A. From (6), α_{x_A}'s can be expressed as a quadratic polynomial of x_A with coefficients from parameters in the RSM model. To estimate these parameters, we can first estimate the α_{x_A}'s, denoted by $\hat{\alpha}_{x_A}$, by least squares and then solve the equations $\hat{\alpha}_{x_A} = \lambda_0 + \lambda_1 x_A + \lambda_{11} x_A^2$, for $x_A = -1,\ 0,\ 1$, to obtain $\hat{\lambda}_0$, $\hat{\lambda}_1$, and $\hat{\lambda}_{11}$. In other words, for x_A^* which is not in $\{-1, 0, 1\}$, we can predict $\alpha_{x_A^*}$ by using $\hat{\lambda}_0 + \hat{\lambda}_1 x_A^* + \hat{\lambda}_{11} (x_A^*)^2$. The same procedure can be applied to β_{x_A}'s and γ_{x_A}'s in (6).

It is then clear why and how the RSM model can be used for prediction. Suppose that we want to predict the value of $E(y)$ at $(x_A^*,\ x_B^*)$, where x_A^* is not included in the experimental plan. From the argument given in Section 2, the NEM cannot be used for prediction at x_A^* because no data are collected at x_A^* for estimating the conditional effects $\alpha_{x_A^*}$, $\beta_{x_A^*}$, and $\gamma_{x_A^*}$. However, the RSM model treats α_{x_A}, β_{x_A}, and γ_{x_A} as continuous (second-order) polynomials over x_A. From this viewpoint and (6), the predicted value of $E(y)$ at $(x_A^*,\ x_B^*)$ is simply $\hat{\alpha}_{x_A^*} + \hat{\beta}_{x_A^*} x_B^* + \hat{\gamma}_{x_A^*} {x_B^*}^2$, where $\hat{\alpha}_{x_A^*}$, $\hat{\beta}_{x_A^*}$, and $\hat{\gamma}_{x_A^*}$ are obtained by substituting x_A^* into the right hand side expressions in (6) with λ's replaced by $\hat{\lambda}$'s. Note that in the prediction procedure using RSM approach, the α_{x_A}, β_{x_A}, and γ_{x_A} are assumed to be *continuous* functions

over x_A and their changes over x_A are assumed to follow the quadratic polynomials in (6). These assumptions explain why prediction is feasible in the RSM, but not in the NEM. In other words, the RSM approach regards the three levels of A as quantitative and utilizes some continuity assumptions on A for prediction. When similar assumptions are imposed on an NEM, prediction using NEM can be feasible.

We will show in Sections 4 and 5 that the RSM model for sliding-level experiment can suffer from severe collinearity between the effects of the slid factor and the effects of its related factors. The RSM model is therefore not a good choice for the purpose of identifying important effects, especially when it is required to perform model selection, such as forward selection or C_p. In these circumstances, we can adopt the following *hybrid strategy* that combines NEM and RSM as follows.

(i) It starts from a NEM, which has better orthogonality between effects in the models.
(ii) After important effects are identified, we can translate the fitted NEM into an RSM model through equations that relate the parameters in the two models (such as (6)). The resulting RSM model can then be used for response prediction.

4. Illustration: a welding experiment

We illustrate the three modeling strategies and compare their results by using data from a welding experiment reported in Chen, Ciscon, and Ratkus [2]. There are eight factors in the experiment: pulse rate (A), weld time (B), cool time (C), hold time (D), squeeze time (E), air pressure (F), current percentage (G), tip size (H). Among them, the pulse rate and the weld time are related factors, i.e., for lower pulse rate, the adequate weld time should be set longer in order to produce weld points with acceptable quality. An 18-run orthogonal array, $OA(18, 2^1 3^7)$, with a slight modification was adopted to study the eight factors. The planning matrix of these factors are given in Table 1 (unfortunately, the units of these factors was not reported). Factors A and H have two levels and other factors have three levels. Note that the column H in Table 1 is obtained by collapsing a three-level factor

Table 1
Planning matrix of the welding experiment

A	B	C	D	E	F	G	H
2	low	6	10	15	50	85	3/8
2	low	12	18	20	55	90	1/4
2	low	18	26	25	60	95	3/8
2	median	6	10	20	55	95	3/8
2	median	12	18	25	60	85	3/8
2	median	18	26	15	50	90	1/4
2	high	6	18	15	60	90	3/8
2	high	12	26	20	50	95	3/8
2	high	18	10	25	55	85	1/4
4	low	6	26	25	55	90	3/8
4	low	12	10	15	60	95	1/4
4	low	18	18	20	50	85	3/8
4	median	6	18	25	50	95	1/4
4	median	12	26	15	55	85	3/8
4	median	18	10	20	60	90	3/8
4	high	6	26	20	60	85	1/4
4	high	12	10	25	50	90	3/8
4	high	18	18	15	55	95	3/8

TABLE 2
Actual settings of pulse rate and weld time

	weld time		
pulse rate	low	median	high
2	32	36	40
4	18	22	26

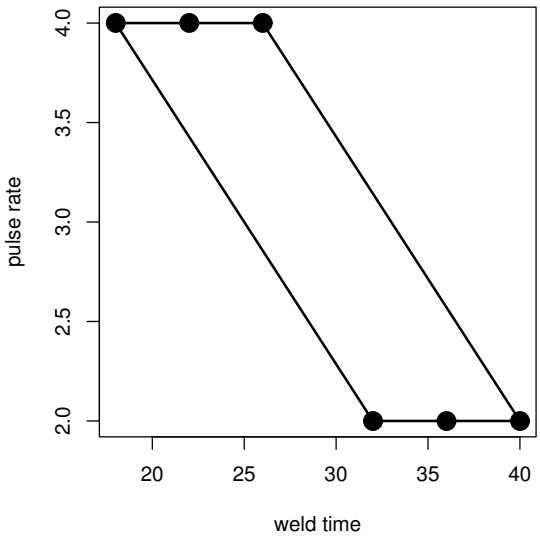


FIG 2. *Adequate experimental region of pulse rate and weld time.*

in the $OA(18, 2^1 3^7)$ to a two-level factor (see Wu and Hamada, [8], Section 7.8). The actual settings of low, median, and high levels in the column B depend on the levels of A as shown in Table 2. We regard the area enclosed by solid lines in Fig 2 as the adequate experimental region, i.e., R_E.

The main difference between the three modeling techniques is reflected in the effect coding of the slid factor B. For the RCRS model, because the R_E is transformed into a square after re-centering and re-scaling, the effect coding of B is the same as in a non-slid factor. Therefore, by applying the linear-quadratic system in [8], Section 5.6, the linear effect of B codes the low, median, and high levels as -1, 0, and 1, respectively, and the quadratic effect of B as 1, -2, 1, respectively. They are shown in the columns labeled by B_l and B_q of Table 3. Note that, although we still call B_l and B_q the main effects of B, they are no longer the main effects of weld time. Instead, the B after RCRS represents a new factor which is a linear combination of weld time and pulse rate. For example, from $B_l = -1$ in Table 3, we can see that the low level of the new factor is the left hand side boundary of R_E in Fig 2 (i.e., the straight line that links the point (weld time, pulse rate)=(32, 2) and the point (weld time, pulse rate)=(18, 4)) and from $B_l = 1$ the high level is the right hand side boundary. For the NEM approach, the effects of B are conditional on the levels of A. For each level of A, the linear-quadratic system is applied to generate the $B_l|A_1$, $B_q|A_1$, $B_l|A_2$, and $B_q|A_2$ as shown in Table 3. For the RSM approach, because the lowest actual setting of B is 18 and the highest actual setting is 40, we code 18 as -1 and 40 as $+1$, and the other settings, 22, 26, 32, and 36, are proportionally coded as $-\frac{7}{11}$, $-\frac{3}{11}$, $\frac{3}{11}$, and $\frac{7}{11}$, respectively. These are shown in the column labeled as x_B in Table 3. The x_B^2 is the componentwise square of x_B.

In the data analysis, we consider the models that contain all main effects of factors C-H and five effects generated from factors A and B. For the RCRS, the

TABLE 3
Effect coding of pulse rate and weld time for the three modeling techniques

factors		RCRS			NEM					RSM		
A	B	A_l	B_l	B_q	A_l	$B_l\|A_1$	$B_q\|A_1$	$B_l\|A_2$	$B_q\|A_2$	x_A	x_B	x_B^2
2	low(32)	−1	−1	1	−1	−1	1	0	0	−1	0.273	0.074
2	median(36)	−1	0	−2	−1	0	−2	0	0	−1	0.636	0.405
2	high(40)	−1	1	1	−1	1	1	0	0	−1	1	1
4	low(18)	1	−1	1	1	0	0	−1	1	1	−1	1
4	median(22)	1	0	−2	1	0	0	0	−2	1	−0.636	0.405
4	high(26)	1	1	1	1	0	0	1	1	1	−0.273	0.074

five effects are A_l, B_l, B_q, A_lB_l, and A_lB_q, where A_lB_l and A_lB_q are interactions generated by the componentwise multiplication of A_l and B_l, and A_l and B_q, respectively. For the NEM, the five effects are A_l, $B_l|A_1$, $B_q|A_1$, $B_l|A_2$, and $B_q|A_2$. For the RSM, the five effects are x_A, x_B, x_B^2, x_Ax_B, and $x_Ax_B^2$, where x_Ax_B and $x_Ax_B^2$ are interactions generated by the componentwise multiplication of x_A and x_B, and x_A and x_B^2, respectively. Although the five effects are coded in different ways for each modeling technique, the vector spaces spanned by any set of the five effects are identical. Consequently, the effects of factors *C-H* will have the same analysis results in the three models. Because of this reason, we only give the analysis results of the five effects generated by A and B, which include their estimated values, t-values, and p-values, under the RCRS, the NEM, and the RSM, in Tables 4, 5, and 6, respectively. From these tables, we have some interesting findings presented in the following.

1. The $B_l|A_1$ and $B_l|A_2$ are the linear effects of B conditional on two different levels of A. In Table 5, we find that the two conditional effects have different magnitudes. When the pulse rate is 2, the weld time has a strong linear effect (significant $B_l|A_1$). When the pulse rate is changed to 4, the linear effect of weld time ($B_l|A_2$) is insignificant. After re-centering and re-scaling, these two effects are transformed into two parameters, B_l and A_lB_l, in Table 4. The B_l represents the average of the two conditional linear effects ($78.96 = ((-23.75) + 181.67)/2$) and the interaction A_lB_l represents the difference between the two conditional linear effects ($-102.71 = ((-23.75) - 181.67)/2$). It is then clear why B_l and A_lB_l are both significant. The same argument can be applied to $B_q|A_1$ and $B_q|A_2$ in Table 5 and B_q and A_lB_q in Table 4. Because $B_q|A_1$ and $B_q|A_2$ has rather similar magnitudes (-27.92 and -41.67), it explains why their difference (i.e., A_lB_q) is insignificant.
2. By comparing x_A in Table 6 and A_l in Tables 4 and 5, we find surprisingly that A_l is significant while x_A is insignificant even though A_l and x_A have the same coding in Table 3. By a further investigation of the correlations between the estimated effects (given in Table 7), it is seen that the insignificance of x_A is caused by the severe collinearity between x_A and x_B. It also results in the other three high correlations in Table 7 because other effects are also defined by x_A and x_B. Note that in the planning matrix in Table 1, all effects in the models based on the RCRS and the NEM are mutually orthogonal. The appearance of severe collinearity will be further explained in Section 5.

   Suppose that severe collinearity is a serious concern but analysis based on RCRS or NEM is not an option to the investigators. A possible choice for reducing the collinearity might be to transform the variables. For example, after replacing weld time in Table 1 by a new variable, pulse rate times weld time, the RSM model will exhibit less correlation between the parameter

TABLE 4
Analysis based on RCRS

	value	t-value	p-value
A_l	−81.04	−6.20	0.00
B_l	78.96	4.93	0.00
B_q	−34.79	−3.76	0.00
$A_l B_l$	−102.71	−6.42	0.00
$A_l B_q$	−6.88	−0.74	0.46

TABLE 5
Analysis based on NEM

	value	t-value	p-value
A_l	−81.04	−6.20	0.00
$B_l \mid A_1$	181.67	8.03	0.00
$B_q \mid A_1$	−27.92	−2.14	0.04
$B_l \mid A_2$	−23.75	−1.05	0.30
$B_q \mid A_2$	−41.67	−3.19	0.00

TABLE 6
Analysis based on RSM

	value	t-value	p-value
x_A	7.72	0.11	0.92
x_B	18.62	0.07	0.95
x_B^2	−789.34	−3.76	0.00
$x_A x_B$	−1287.06	−4.76	0.00
$x_A x_B^2$	−155.98	−0.74	0.46

TABLE 7
Correlation matrix of the estimated effects under the RSM model

	x_B	x_B^2	$x_A x_B$	$x_A x_B^2$
x_A	0.96	0.00	0.00	0.91
x_B		0.00	0.00	0.99
x_B^2			0.99	0.00
$x_A x_B$				0.00

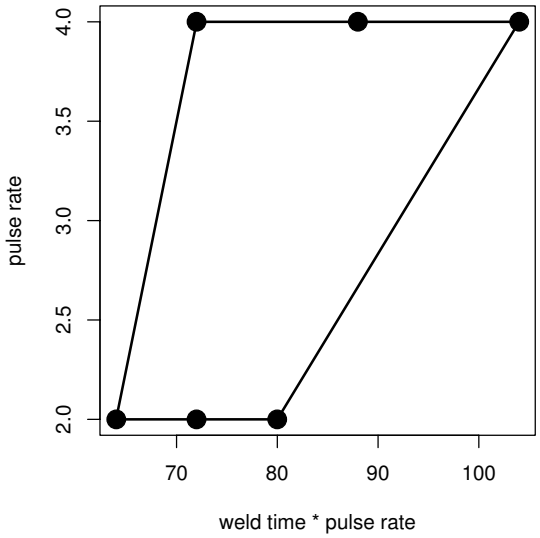


FIG 3. *Transformed experimental region.*

estimates because the transformed experimental region (given in Fig 3) is

more similar to a rectangle than the original experimental region (as shown in Fig 2).

3. From Table 3, we can understand that for the main effects of the slid factor (which include B_l and B_q in RCRS, $B_l|A_i$ and $B_q|A_i$ in NEM, and x_B and x_B^2 in RSM), the coding based on RCRS can best preserve orthogonality property in a planning matrix, followed by the NEM, and RSM being the worst.

5. Relationship between RCRS and RSM models

To explain the relationship between the RCRS and RSM models, consider an RCRS model for the nine-run experiment that contains all main effects and a linear-by-linear interaction as follows:

$$
\begin{aligned}
f(x_A, x_B) \approx \eta_0 + \eta_1 x_A + \eta_{11} x_A^2 + \eta_2 \left[\frac{x_B - m_B(x_A)}{l_B(x_A)}\right] \\
+ \eta_{22} \left[\frac{x_B - m_B(x_A)}{l_B(x_A)}\right]^2 + \eta_{12} x_A \left[\frac{x_B - m_B(x_A)}{l_B(x_A)}\right],
\end{aligned} \tag{7}
$$

where $m_B(x_A)$ and $l_B(x_A)$ are the center and range, respectively, of the experimental region chosen for B when A is conditioned on x_A. For simplicity, assume that m_B is a linear function of x_A, i.e., $m_B(x_A) = s + t\ x_A$, and l_B is a constant, i.e., $l_B(x_A) = r$ (same shape as Fig 2). By substituting them into (7) and expanding (7) in a polynomial form, we obtain an RSM model, consisting of the factorial effects x_A, x_A^2, x_B, x_B^2, and $x_A x_B$, as follows:

$$
\begin{aligned}
f(x_A, x_B) \approx & \left[\eta_0 + (s^2/r^2)\eta_{22} - (s/r)\eta_2\right] \\
& + \left[\eta_1 - (t/r)\eta_2 - (s/r)\eta_{12} + (2st/r^2)\eta_{22}\right] x_A \\
& + \left[\eta_{11} + (t^2/r^2)\eta_{22} - (t/r)\eta_{12}\right] x_A^2 + \left[(1/r)\eta_2 - (2s/r^2)\eta_{22}\right] x_B \\
& + \left[(1/r^2)\eta_{22}\right] x_B^2 + \left[(1/r)\eta_{12} - (2t/r^2)\eta_{22}\right] x_A x_B.
\end{aligned} \tag{8}
$$

Note that in (8), the parameters of factorial effects are functions of η's and r, s, and t. The η's represent the relationship between factors and response in the RCRS model, and r, s, and t characterize the shape of the irregular experimental region. The shape has been eliminated in the RCRS model after applying the transformation $\frac{x_B - m_B(x_A)}{l_B(x_A)}$ on B. However, m_B and l_B still affect the polynomial terms of the RSM model in (8). This example shows that an RSM model for sliding-level experiments contains two components: a description of the relationship between factors and response, and a description of the irregular shape of the experimental region. The two components are intertwined and undistinguishable in the parameters of an RSM model. On the other hand, a fitted model based on RCRS only contains information on the first component because the irregularity of shape has been eliminated after re-centering and re-scaling. This observation is supported by the appearance of strong collinearity between x_A and x_B in Table 7. Note that after RCRS, the main effects of A and B (in Table 4) are orthogonal. However, in the RSM model such strong collinearity inevitably appears because: (i) the parameters in the RSM model are influenced by the irregular shape of experimental region, and (ii) the irregular shape (i.e., R_E in Fig 2) reflects the fact that B is smaller when A is larger.

In general, the shape of the chosen experimental region can be arbitrary, and m_B and l_B can have more complicated forms than what was assumed above. However,

similar remarks and conclusions are still applicable.

Suppose that the relationship between mean response and factors A and B satisfies (1). When the m_B and l_B in (7) are appropriately chosen so that the interaction elimination after RCRS is achieved, the η_{12} in (7) becomes zero. In this case, the RCRS model in (7) does not nominally contain an interaction effect, but an interaction (i.e., $x_A x_B$) is still present in the RSM model (8). The apparent discrepancy lies in the different approaches they take to handle the irregular shape of the experimental region. This observation partially supports the interaction elimination rationale in RCRS from a different perspective. Because the fitted model after RCRS does not properly take into account the irregular shape of the experimental region, it can, in most cases, utilize fewer effects than an RSM model to achieve a comparable coefficient of determination (i.e., R^2). Some interaction effects (such as the $x_A x_B$ in the case) are not required for the RCRS model.

6. Summary

For the purpose of response prediction for sliding-level experiments, we point out the shortcomings of two existing approaches, RCRS and NEM, when the related factors are quantitative. An alternative analysis strategy is proposed based on the response surface modeling, in which the response prediction can be implemented in a straightforward manner. Through the comparisons of the three strategies, we present several interesting conclusions, which lead to better understanding of the concepts, properties, limitations, and implicit assumptions behind each strategy. None of the three methods dominates the others in every aspects. The best strategy for the investigators depends on the information they have about the irregular region and the objectives of the experiment. Although we do not discuss the design issues in this article, the choice of the modeling strategy will influence the choice of the best design. This and other issues in modeling deserve further study.

Acknowledgments

S.-W. Cheng's research was supported by the National Science Council of Taiwan, ROC. C.F.J. Wu's research was supported by the NSF. L. Huwang's research was supported by the National Science Council of Taiwan, ROC, under the grant No. NSC-94-2118-M-007-003. The authors are grateful to the referee for valuable comments.

References

[1] Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, New York.

[2] Chen, T. Y. W., Ciscon, K. J. and Ratkus, A. C. (1984). ITT Bell and Gossett Taguchi experiment: Motor support bracket spot welding. *Second Symposium on Taguchi Methods*. American Supplier Institute, Romulus, MI, pp. 70–77.

[3] Hamada, M. and Wu, C. F. J. (1995). The treatment of related experimental factors by sliding levels. *Journal of Quality Technology* **27** 45–55.

[4] Li, W., Cheng, S.-W., Hu, S. J. and Shriver, J. (2001). Statistical investigation on resistance spot welding quality using a two-stage, sliding-level experiment. *ASME Journal of Manufacturing Science and Engineering* **123** 513–520.

[5] Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley & Sons, New York.

[6] Phadke, M. S., Kackar, R. N., Speeney, D. V., and Grieco, M. J. (1983). Off-line quality control in integrated circuit fabrication using experimental design. *The Bell System Technical Journal* **62** 1273–1309.

[7] Taguchi, G. (1987). *System of Experimental Design*. Unipub/ Kraus International Publications, White Plains, NY.

[8] Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, New York.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 257–271

DOI: 10.1214/074921706000001094

Price systems for markets with transaction costs and control problems for some finance problems

Tzuu-Shuh Chiang[1], Shang-Yuan Shiu[2] and Shuenn-Jyi Sheu[1,*]

Academia Sinica and University of Utah

Abstract: In a market with transaction costs, the price of a derivative can be expressed in terms of (preconsistent) price systems (after Kusuoka (1995)). In this paper, we consider a market with binomial model for stock price and discuss how to generate the price systems. From this, the price formula of a derivative can be reformulated as a stochastic control problem. Then the dynamic programming approach can be used to calculate the price. We also discuss optimization of expected utility using price systems.

1. Introduction

Duality approach is frequently used for financial problems in incomplete markets. This approach can also be applied to markets with transaction costs. In [12], a discrete market with transaction costs is considered. In the market studied there is a stock and a bond that we can trade. Let $\lambda_1, \lambda_0 > 0$ be the proportional costs for selling and buying the stock. Then the replication cost at time 0 for a portfolio $Y = (Y^0, Y^1)$ at time T is given by

$$\pi^*(Y) = \sup\{E[Y^0\rho^0 + Y^1\rho^1]\}. \tag{1.1}$$

The supremum is taken over (ρ^0, ρ^1) ((preconsistent) price systems) which depend on λ_0, λ_1. This will be described in details below.

A similar result for diffusion models is given in [3].

Our interest is to use price systems to calculate the price of a derivative and find optimal strategy for hedging problem. We will also discuss the use of price systems to study portfolio optimization problem. There is a similarity between these problems that they can be reformualted as optimization problems. We shall consider binomial model (it can also be extended to multinomial model) and find a dynamics to generate the price systems (ρ^0, ρ^1). A price system becomes a controlled process. The optimization problems become stochastic control problems. Then dynamic programming approach can be used.

The paper is organized as follows. In Section 2, we give notations and give the framework. In Section 3, we describe price systems and give a price formula for derivatives in terms of price systems. In Section 4, we discuss the optimization of

*Supported by the grant NSC 93-2115-M-001-010.

[1]Institute of Mathematics, Academia Sinica, Nankang, Taipei, Taiwan, e-mail: matsch@math.sinica.edu.tw; sheusj@math.sinica.edu.tw

[2]Department of Mathematics, University of Utah, 155 South 1400 East, Salt Lake City, UT 84112-0090, USA.

AMS 2000 subject classifications: 60K35, 60K35.

Keywords and phrases: dynamic programming, duality method, price system, pricing derivatives, portfolio optimization, stochastic control, transaction cost.

expected utility using price systems. In Sections 5, 6 and 7, we consider binomial models. We present a dynamics to generate the price systems. We reformulate some finance problems as stochastic control problems. Then we use dynamic programming to calculate the value functions.

2. Finite market with one stock

The framework can be described as follows.

We consider $(\Omega, \mathcal{F}, P)$ a finite probability space and $\{\mathcal{F}_k\}$ a filtration. Let

$$P^0(k;\omega), P^1(k;\omega)$$

be the prices for bond and stock. Then P^0, P^1 are adapted to $\{\mathcal{F}_k\}$. Define

$$\hat{P}(k;\omega) = P^1(k,\omega)/P^0(k;\omega),$$

the discounted price.

A trading strategy is given by $\{I(k;\omega)\}_{k=0}^T$, a stochastic process adapted to $\{\mathcal{F}_k\}$. $I(k;\omega)$ is the number of shares that the stock is bought or sold,

$$\begin{aligned} I(k;\omega) \geq 0, &\quad \text{buy stock at } k, \\ I(k;\omega) < 0, &\quad \text{sell stock at } k. \end{aligned}$$

The portfolio values for $\{I(k;\omega)\}_{k=0}^T$ with $x = (x_0, x_1)$ are given by,

$$\begin{aligned} X^0(k;x,I) &= x^0 - \sum_{\ell=0}^{k} h(I(\ell))\hat{P}(\ell) \\ X^1(k;x,I) &= x^1 + \sum_{\ell=0}^{k} I(\ell). \end{aligned}$$

Here

$$h(z) = \begin{cases} (1+\lambda_0)z, & z > 0 \\ (1-\lambda_1)z, & z \leq 0, \end{cases}$$

where $\lambda_1, \lambda_0 > 0$ are the proportional costs for selling and buying the stock, respectively.

We are interested in the following finance problems.

Pricing derivative: Let $Y = (Y^0, Y^1)$ be $\mathcal{F}_T$ measurable. We define $\pi^*(Y)$ the minimum of $x_0 P^0(0)$ such that for some I,

$$Y^0 \leq X^0(T;(x^0,0),I),\ Y^1 \leq X^1(T;(x^0,0),I).$$

We say $\pi^*(Y)$ is the price of $Y = (Y^0, Y^1)$. The problem is to calculate $\pi^*(Y)$. Another important problem is to obtain a strategy $I(\cdot)$ such that for $x^0 = \pi^*(Y)$,

$$Y^0 \leq X^0(T;(x^0,0),I),\ Y^1 \leq X^1(T;(x^0,0),I).$$

For the later use, we also define $\pi^*(Y;x_1)$ the minimum of $x_0 P^0(0)$ such that for some I,

$$Y^0 \leq X^0(T;(x^0,x_1),I),\ Y^1 \leq X^1(T;(x^0,x_1),I).$$

Then $\pi^*(Y;x_1) = \pi^*(\tilde{Y})$, where $\tilde{Y}^0 = Y^0, \tilde{Y}^1 = Y^1 - x_1$.

Optimizing expected utility: Let U be a utility function. Let (x^0, x^1) be given such that

$$x^0 P^0(0) - h(-x^1)P^1(0) > 0.$$

$V(x^0, x^1)$ is the maximum of

$$E[U(X^0(T;(x^0,x^1),I)P^0(T) - h(-X^1(T;(x^0,x^1),I))P^1(T))],$$

where $I(\cdot)$ is an admissible strategy: $k = 0, 1, 2, \ldots, T$,

$$X^0(k;(x^0,x^1),I)P^0(k) - h(-X^1(k;(x^0,x^1),I))P^1(k) \geq 0.$$

We want to calculate $V(x^0, x^1)$ and find a strategy I that attains the maximum.

3. Price systems and a price formula

Definition. We say that (ρ^0, ρ^1) is a price system if ρ^0, ρ^1 are positive random variables such that

(a) $E[\rho^0] = P^0(0)$;
(b) Define

$$R(k;\omega) = \frac{\rho^1(k;\omega)}{\rho^0(k;\omega)} \frac{1}{\hat{P}(k;\omega)},$$

$$\rho^0(k;\omega) = E[\rho^0|\mathcal{F}_k],\ \rho^1(k;\omega) = E[\rho^1|\mathcal{F}_k].$$

Then

$$(1-\lambda_1) \leq R(k;\omega) \leq (1+\lambda_0),\ k = 0, 1, 2, \ldots, T.$$

We denote $\mathcal{P}(\lambda_0, \lambda_1)$ the family of price systems.

Remark. Assume there is an equivalent martingale measure Q. Then $\mathcal{P}(\lambda_0, \lambda_1) \neq \emptyset$. In fact, define

$$\rho^0 = \frac{dQ}{dP} P^0(0)$$

$$\rho^1 = \rho^0 \hat{P}(T).$$

Then

$$\rho^0(k;\omega) = \frac{dQ}{dP}|_{\mathcal{F}_k} P^0(0)$$

$$\rho^1(k;\omega) = \rho^0(k;\omega)\hat{P}(k;\omega).$$

We can show that (ρ^0, ρ^1) is a price system.

On the other hand, in the case $\lambda_0 = \lambda_1 = 0$, (ρ^0, ρ^1) is a price system if and only if

$$\frac{dQ}{dP} = \rho^0(T)/P^0(0)$$

defines an equivalent martingale measure.

Theorem 1 ([12]). *Assume $\mathcal{P}(\lambda_0, \lambda_1) \neq \emptyset$. Then*

$$\pi^*(Y) = \sup_{\mathcal{P}(\lambda_0,\lambda_1)} E[Y^0\rho^0 + Y^1\rho^1]. \tag{3.1}$$

Remark. If $\lambda_0 = \lambda_1 = 0$, then the above is

$$\pi^*(Y) = \sup_{\rho} E[\rho H \frac{P^0(0)}{P^0(T)}]$$

$$H = Y^0 P^0(T) + Y^1 P^1(T).$$

Here

$$\frac{dQ}{dP} = \rho$$

defines an equivalent martingale measure.

A similar result for diffusion models is given in [3].

4. Price system and optimal expected utility

In the following, we assume $P^0(k) = 1$ for all k.

Let U be a strictly increasing utility function. Define

$$U^*(y) = \sup\{U(x) - xy; x \geq 0\}.$$

Define

$$V^*(\xi, x_1) = \inf\{E[U^*(\xi\rho^0(T))] + x_1\xi E[\rho^1(T)]\}$$

Theorem 2. *We have*

$$V(x_0, x_1) \leq \inf_{\xi>0}\{V^*(\xi, x_1) + x_0\xi\}$$

This is the same as

$$V(x_0, x_1) \leq \inf_{\xi>0,\rho^0,\rho^1}\{E[U^*(\xi\rho^0(T))] + x_1\xi E[\rho^1(T)] + x_0\xi\}. \tag{4.1}$$

Assume there is $\hat{\xi}, \hat{\rho}^0, \hat{\rho}^1$ that attains the infimum. Then the above equality holds. Moreover, there is an optimal strategy $\hat{I}$ for the portfolio optimization problem satisfying the following properties.

(a) $X^0(T; (x_0, x_1), \hat{I}) = -U^{*'}(\hat{\xi}\hat{\rho}^0(T))$,
$X^1(T; (x_0, x_1), \hat{I}) = 0$.

(b) $\hat{R}(l) = 1 + \lambda_0$ *if* $\hat{I}(l) > 0$,
$\hat{R}(l) = 1 - \lambda_1$ *if* $\hat{I}(l) < 0$.

Here $U^{*'}(\xi)$ denotes the derivative of $U^*(\xi)$.

Proof. Let I be a strategy.

$$\begin{aligned}
&U(X^0(T; x, I) - h(-X^1(T; x, I))P^1(T)) \\
&\quad \leq U^*(\xi\rho^0(T)) + \xi\rho^0(T)(X^0(T; x, I) - h(-X^1(T; x, I))P^1(T)) \\
&\quad = U^*(\xi\rho^0(T)) + \xi(X^0(T; x, I)\rho^0(T) - h(-X^1(T; x, I))P^1(T)\rho^0(T)) \\
&\quad \leq U^*(\xi\rho^0(T)) + \xi(X^0(T; x, I)\rho^0(T) + R(T)X^1(T; x, I)P^1(T)\rho^0(T)) \\
&\quad = U^*(\xi\rho^0(T)) + \xi(X^0(T; x, I)\rho^0(T) + \rho^1(T)X^1(T; x, I)).
\end{aligned} \tag{4.2}$$

$$\begin{aligned}
&X^0(T; x, I)\rho^0(T) + X^1(T; x, I)\rho^1(T) \\
&\quad = x_0\rho^0(T) + x_1\rho^1(T) + (-\sum_{l=0}^{T} h(I(l))P^1(l)\rho^0(T) \\
&\qquad + \sum_{l=0}^{T} I(l)\rho^1(T)).
\end{aligned} \tag{4.3}$$

$$
\begin{aligned}
&E[(-\sum_{l=0}^{T} h(I(l))P^1(l)\rho^0(T) + \sum_{l=0}^{T} I(l)\rho^1(T))] \\
&= \sum_{l=0}^{T} E[-h(I(l))P^1(l)\rho^0(l) + I(l)\rho^1(l)] \\
&= \sum_{l=0}^{T} E[(-h(I(l)) + R(l)I(l))P^1(l)\rho^0(l)] \\
&\leq 0.
\end{aligned} \tag{4.4}
$$

Then we can deduce

$$
\begin{aligned}
&E[U(X^0(T;x,I) - h(-X^1(T;x,I))P^1(T))] \\
&\quad \leq E[U^*(\xi\rho^0(T))] + \xi x_1 E[\rho^1(T)] + \xi x_0.
\end{aligned} \tag{4.5}
$$

This is true for all ρ^0, ρ^1. The first result follows.

Assume $\hat{\xi}, \hat{\rho}^0, \hat{\rho}^1$ attains infimum in (4.1). Then

$$
E[U^{*'}(\hat{\xi}\hat{\rho}^0(T))\hat{\rho}^0(T)] + x_1 E[\hat{\rho}^1(T)] + x_0 = 0. \tag{4.6}
$$

On the other hand, take any (ρ^0, ρ^1) and $0 < \alpha < 1$, we have

$$
E[U^*(\hat{\xi}(\alpha\rho^0(T) + (1-\alpha)\hat{\rho}^0(T)))] + x_1\hat{\xi}E[\alpha\rho^1(T) + (1-\alpha)\hat{\rho}^1(T)] + x_0\hat{\xi}
$$

takes minimum at $\alpha = 0$. We have

$$
\begin{aligned}
&E[U^{*'}(\hat{\xi}\hat{\rho}^0(T))(\rho^0(T) - \hat{\rho}^0(T))] \\
&\qquad + x_1\hat{\xi}E[\rho^1(T) - \hat{\rho}^1(T)] \geq 0.
\end{aligned} \tag{4.7}
$$

Take

$$
\hat{Y}^0 = -U^{*'}(\hat{\xi}\hat{\rho}^0(T)),\ \hat{Y}^1 = 0.
$$

(4.7) implies

$$
\pi^*(\hat{Y}; x_1) = E[-U^{*'}(\hat{\xi}\hat{\rho}^0(T))\hat{\rho}^0(T)] - x_1\hat{\xi}E[\hat{\rho}^1(T)] = x_0.
$$

Here we use (4.6) and Theorem 1.

By the definition of $\pi^*(\hat{Y}; x_1)$, there is a strategy $\hat{I}$ such that

$$
\begin{aligned}
x_0 - \sum_{l=0}^{T} h(\hat{I}(l))P^1(l) &\geq \hat{Y}^0, \\
x_1 + \sum_{l=0}^{T} \hat{I}(l) &\geq \hat{Y}^1.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&X^0(T;(x_0,x_1),\hat{I}) \geq -U^{*'}(\hat{\xi}\hat{\rho}^0(T)), \\
&X^1(T;(x_0,x_1),\hat{I}) \geq 0.
\end{aligned} \tag{4.8}
$$

$$
\begin{aligned}
&U(X^0(T;x,\hat{I}) - h(-X^1(T;x,\hat{I}))P^1(T)) \\
&\quad \geq U(-U^{*'}(\hat{\xi}\hat{\rho}^0(T))) \\
&\quad = U^*(\hat{\xi}\hat{\rho}^0(T)) - \hat{\xi}\hat{\rho}^0(T)U^{*'}(\hat{\xi}\hat{\rho}^0(T)).
\end{aligned} \tag{4.9}
$$

Then

$$\begin{aligned} & E[U(X^0(T;x,\hat{I}) - h(-X^1(T;x,\hat{I}))P^1(T))] \\ & \quad \geq E[U^*(\hat{\xi}\hat{\rho}^0(T))] - \hat{\xi}E[\hat{\rho}^0(T)U^{*'}(\hat{\xi}\hat{\rho}^0(T)))] \\ & \quad = E[U^*(\hat{\xi}\hat{\rho}^0(T))] + \hat{\xi}(x_1 E[\hat{\rho}^1(T)] + x_0) \\ & \quad \geq V(x_0, x_1). \end{aligned} \tag{4.10}$$

Therefore, by the definition of $V(x_0, x_1)$, the inequalities become equalities in the above relation. We see (a) follows from the equalities in (4.8),(4.9) and (4.10). On the other hand, (4.2),(4.3), (4.4) and (4.5) also become equalities for $I = \hat{I}$, then (b) follows. This completes the proof. □

5. Binomial model and price systems

We take $P_k^0 = 1$ for all k. $0 < d < 1 < u$, $\lambda_0, \lambda_1 > 0$.

The sample space is given by

$$\Omega = \{(a_1, a_2, \ldots, a_T); a_i \in \{u, d\}\}.$$

For $\omega = (a_1, a_2, \ldots, a_T) \in \Omega$, denote

$$\omega^k = (a_1, a_2, \ldots, a_k).$$

The price of stock is

$$P_k^1(\omega) = P_0^1 a_1 a_2 \cdots a_k.$$

We also write $P_k^1(\omega) = P_k^1(\omega^k)$.

$\mathcal{F}_k$ is the σ-algebra generated by $P_t^1, t \leq k$. A function defined on Ω measurable w.r.t. $\mathcal{F}_k$ is given by $f(\omega^k)$.

For $\omega = (a_1, a_2, \ldots, a_T) \in \Omega$, the probability is given by

$$P(\{\omega\}) = p^m (1-p)^{T-m},$$

where m is the number of k such that $a_k = u$, $0 < p < 1$.

$\rho^0(k), \rho^1(k)$ are given by

$$\rho^0(k) = E[\rho^0 | \mathcal{F}_k], \ \rho^1(k) = E[\rho^1 | \mathcal{F}_k].$$

We have the characterization of $\rho^0(k), \rho^1(k)$:

$$\begin{aligned} \rho^0(k, \omega^k) &= p\rho^0(k+1, (\omega^k, u)) + (1-p)\rho^0(k+1, (\omega^k, d)), \\ \rho^1(k, \omega^k) &= p\rho^1(k+1, (\omega^k, u)) + (1-p)\rho^1(k+1, (\omega^k, d)). \end{aligned} \tag{$PS1$}$$

$$(1-\lambda_1)P_k^1 \leq \frac{\rho^1(k)}{\rho^0(k)} \leq (1+\lambda_0)P_k^1, k = 1, 2, 3, \ldots, T. \tag{$PS2$}$$

It is convenient to consider

$$A(k) = \frac{\rho^1(k)}{\rho^0(k)}, \ k = 0, 1, \ldots, T.$$

We can now describe the price systems in a binomial market. We omit the easy proof.

Theorem 3 (Binomial model). *Let* $\omega = (a_1, a_2, \ldots, a_T) \in \Omega$. *Given* A_0 *a positive constant such that*

$$(1 - \lambda_1)P_0^1 \leq A_0 \leq (1 + \lambda_0)P_0^1.$$

Denote $\rho^0(0) = 1, \rho^1(0) = \rho^0(0)A_0$. *Take positive constants* A^u, A^d *such that*

$$\min\{A^u, A^d\} < A_0 < \max\{A^u, A^d\}$$

and

$$(1 - \lambda_1)P_0^1 u \leq A^u \leq (1 + \lambda_0)P_0^1 u,$$

$$(1 - \lambda_1)P_0^1 d \leq A^d \leq (1 + \lambda_0)P_0^1 d.$$

If $a_1 = u$

$$\rho^0(1) = \frac{1}{p}\frac{A_0 - A^d}{A^u - A^d},$$

$$A_1 = A^u.$$

If $a_1 = d$

$$\rho^0(1) = \frac{1}{1-p}\frac{A^u - A_0}{A^u - A^d},$$

$$A_1 = A^d.$$

Define

$$\rho^1(1) = \rho^0(1)A_1.$$

Assume we have defined $A_0, A_1, \ldots, A_k$ *and*

$$\rho^0(1), \rho^0(2), \ldots, \rho^0(k), \rho^1(1), \rho^1(2), \ldots, \rho^1(k).$$

Take A^u, A^d *measurable w.r.t.* $\mathcal{F}_k$ *such that*

$$\min\{A^u, A^d\} < A_k < \max\{A^u, A^d\}$$

and

$$(1 - \lambda_1)P_k^1 u \leq A^u \leq (1 + \lambda_0)P_k^1 u,$$

$$(1 - \lambda_1)P_k^1 d \leq A^d \leq (1 + \lambda_0)P_k^1 d.$$

If $a_{k+1} = u$,

$$\rho^0(k+1) = \rho^0(k)\frac{1}{p}\frac{A_k - A^d}{A^u - A^d},$$

$$\rho^1(k+1) = \rho^0(k+1)A^u;$$

if $a_{k+1} = d$,

$$\rho^0(k+1) = \frac{1}{1-p}\frac{A^u - A_k}{A^u - A^d},$$

$$\rho^1(k+1) = \rho^0(k+1)A^d.$$

Then $\rho^0(k), \rho^1(k)$ *satisfy (PS1) and (PS2).*

6. Binomial model: control problems for pricing derivatives

Assume $Y = (Y^0, Y^1)$ is given by

$$Y^0 = Y^0(P_T^1),\ Y^1 = Y^1(P_T^1).$$

Then the price $\pi^*(Y)$ is given by

$$\pi^*(Y) = \sup_{\rho^0(T),\rho^1(T)} E[\rho^0(T)Y^0(P_T^1) + \rho^1(T)Y^1(P_T^1)].$$

This can be rewritten as

$$\pi^*(Y) = \sup_{\rho^0(T),\rho^1(T)} E[\rho^0(T)(Y^0(P_T^1) + A_T Y^1(P_T^1))].$$

with A_k and $\rho^0(k)$ described in Theorem 3. This is viewed as a stochastic control problem. The state variables are given by $P_k^1, A_k, \rho^0(k)$ and the control variables are A_k^u, A_k^d.

The dynamical programming can be described as follows.

For $S > 0$ and A satisfying

$$(1-\lambda_1)S \le A \le (1+\lambda_0)S,$$

define

$$W_k(S,A) = \sup E[\frac{\rho^0(T)}{\rho^0(k)}(Y^0(P_T^1) + A_T Y^1(P_T^1))|P_k^1 = S, A_k = A]$$

Then

$$\pi^*(Y) = \sup_{(1-\lambda_1)S \le A \le (1+\lambda_0)S} W_0(S,A)$$

for $P_0^1 = S$. And for $0 \le k < l \le T$,

$$W_k(S,A) = \sup E[\frac{\rho^0(l)}{\rho^0(k)} W_l(P_l^1, A_l)|P_k^1 = S, A_k = A]$$

It follows a recursive scheme backward in time.

(D1) $W_T(S,A) = Y^0(S) + AY^1(S)$;

(D2) For $(1-\lambda_1)S \le A \le (1+\lambda_0)S$,

$$W_k(S,A) = \sup\{\frac{A-A^d}{A^u-A^d} W_{k+1}(Su, A^u) + \frac{A^u-A}{A^u-A^d} W_{k+1}(Sd, A^d)\},$$

the maximization is taken over

$$\min\{A^u, A^d\} < A < \max\{A^u, A^d\}$$

$$(1-\lambda_1)Su \le A^u \le (1+\lambda_0)Su,$$

$$(1-\lambda_1)Sd \le A^d \le (1+\lambda_0)Sd.$$

(D3) For $P_0^1 = S$,

$$\pi^*(Y) = \sup_{(1-\lambda_1)S \le A \le (1+\lambda_0)S} \{W_0(S,A)\}.$$

It can be restated as follows.

Theorem 4. *We have*

$$W_k(S,A) = \sup\{\alpha W_{k+1}(Su, A^u) + (1-\alpha)W_{k+1}(Sd, A^d)\},$$

the maximization is taken over

$$0 < \alpha < 1,$$
$$\alpha A^u + (1-\alpha)A^d = A,$$

and

$$(1-\lambda_1)Su \le A^u \le (1+\lambda_0)Su,$$
$$(1-\lambda_1)Sd \le A^d \le (1+\lambda_0)Sd.$$

$W_k(S,A)$ is piecewise linear in A for all $S > 0$ and $k = 0, 1, \ldots, T$.

The main questions consist of the following. How to calculate $W_k(S,A)$? How to obtain an optimal strategy to super hedge Y from $W_k(S,A)$? Some answers can be found in [6].

7. Binomial model: optimizing expected utility and control problem

We take

$$U(x) = \frac{1}{\gamma}x^\gamma, 0 < \gamma < 1.$$

Then

$$U^*(\xi) = -\frac{1}{\mu}\xi^\mu,$$
$$\mu = \frac{\gamma}{\gamma - 1}.$$

Then

$$V(x_0, x_1) = \inf_{\xi > 0}\{V^*(\xi, x_1) + x_0\xi\}, \tag{7.1}$$

$$V^*(\xi, x_1) = \inf\{-\frac{1}{\mu}\xi^\mu E[(\rho^0(T))^\mu] + \xi x_1 E[\rho^1(T)]\}.$$

We shall consider

$$-\frac{1}{\mu}\xi^\mu E[(\rho^0(T))^\mu] + \xi x_1 E[\rho^1(T)]$$

conditioning on $P^1(0) = S, A(0) = A$. This is equal to

$$-\frac{1}{\mu}\xi^\mu E[(\rho^0(T))^\mu] + \xi x_1 A.$$

We consider

$$V_k(S,A) = \inf E[(\frac{\rho^0(T)}{\rho^0(k)})^\mu | \mathcal{F}_k].$$

The follwoing is an iterative scheme to calculate $V_k(S,A), k = 0, 1, \ldots$.

(PD1) $V_T(S,A) = 1$,

(PD2) $k = 0, 1, 2, \ldots$,

$$V_k(S,A) = \inf\{p^{1-\mu}(\frac{A - A^d}{A^u - A^d})^\mu V_{k+1}(Su, A^u) + (1-p)^{1-\mu}(\frac{A^u - A}{A^u - A^d})^\mu V_{k+1}(Sd, A^d)\} \tag{7.2}$$

The inf is taken over the A^d, A^u satisfying

$$\min\{A^u, A^d\} < A < \max\{A^u, A^d\}$$

$$\begin{aligned}(1-\lambda_1)Su \le A^u \le (1+\lambda_0)Su,\\ (1-\lambda_1)Sd \le A^d \le (1+\lambda_)Sd.\end{aligned} \tag{7.3}$$

(7.2) can be reformulated as follows.

$$V_k(S,A) = \inf\{p^{1-\mu}\alpha^\mu V_{k+1}(Su, A^u) + (1-p)^{1-\mu}(1-\alpha)^\mu V_{k+1}(Sd, A^d)\} \tag{7.2}'$$

where $0 \le \alpha \le 1$ and (7.3) and (7.4) hold,

$$\alpha A^u + (1-\alpha)A^d = A. \tag{7.4}$$

We consider $V_{T-1}(S,A)$:

$$V_{T-1}(S,A) = \inf\{p^{1-\mu}\alpha^\mu + (1-p)^{1-\mu}(1-\alpha)^\mu\},$$

where (7.3), (7.4) hold. Denote $\hat{V}_{T-1}(A) = V_{T-1}(S, SA)$. For

$$(1-\lambda_1) \le A \le (1+\lambda_0),$$

$$\hat{V}_{T-1}(A) = \inf\{p^{1-\mu}\alpha^\mu + (1-p)^{1-\mu}(1-\alpha)^\mu\},$$

there are A^u, A^d such that

$$\begin{aligned}(1-\lambda_1) \le A^u \le (1+\lambda_0),\\ (1-\lambda_1) \le A^d \le (1+\lambda_0).\end{aligned} \tag{7.3}'$$

$$\alpha u A^u + (1-\alpha)dA^d = A, \tag{7.4}'$$

In general,

$$\hat{V}_k(A) = V_k(S, SA),\ (1-\lambda_1) \le A \le (1+\lambda_0).$$

Then

$$\hat{V}_k(A) = \inf\{p^{1-\mu}\alpha^\mu \hat{V}_{k+1}(A^u) + (1-p)^{1-\mu}(1-\alpha)^\mu \hat{V}_{k+1}(A^d)\},$$

$0 \le \alpha \le 1$ satisfies (7.3)′, (7.4)′. From

$$\hat{V}_k(A),\ (1-\lambda_1) \le A \le (1+\lambda_0),$$

we have

$$V_k(S,A) = \hat{V}_k(\frac{A}{S}),\ (1-\lambda_1)S \le A \le (1+\lambda_0)S.$$

Theorem 5. *Assume $x_0 - h(-x_1)S > 0$. Then for $P^1(0) = S$,*

$$V(x_0, x_1) = \frac{1}{\gamma} inf\{(x_0 + x_1 SR)^\gamma (\hat{V}_0(R))^{1-\gamma}\}$$

the infimum is taken over $(1-\lambda_1) \le R \le (1+\lambda_0)$.

In particular, if $1 < pu + (1-p)d$ and $x_1 \le 0$, then

$$V(x_0, x_1) = \frac{1}{\gamma}(x_0 + x_1 S(1+\lambda_0))^\gamma (\hat{V}_0(1+\lambda_0))^{1-\gamma}.$$

If $1 > pu + (1-p)d$ and $x_1 \ge 0$, then

$$V(x_0, x_1) = \frac{1}{\gamma}(x_0 + x_1 S(1-\lambda_1))^\gamma (\hat{V}_0(1-\lambda_1))^{1-\gamma}.$$

Proof. By (7.1)

$$V(x_0, x_1) = \inf\{-\frac{1}{\mu}\xi^{\mu}\hat{V}_0(R) + \xi x_1 SR + \xi x_0\},$$

the inf is taken over $\xi > 0, (1-\lambda_1) \le R \le (1+\lambda_0)$.

$$\hat{\xi} = (\frac{1}{\hat{V}_0(R)}(x_0 + x_1 SR))^{\frac{1}{\mu-1}}$$

takes minimum. The rest follows from this and Theorem 6 below. □

Theorem 6. *Assume* $1 < pu + (1-p)d$. *Then for*

$$(1-\lambda_1)(pu + (1-p)d)^{T-k} \le A \le (1+\lambda_0),$$

$\hat{V}_k(A) = 1$. *For other* A, $\hat{V}_k(A) > 1$ *and is decreasing in* A.
Assume $pu + (1-p)d < 1$. *Then for*

$$(1-\lambda_1)S \le A \le (1+\lambda_0)S(pu + (1-p)d)^{T-k},$$

$\hat{V}_k(A) = 1$. *For other* A, $\hat{V}_k(A) > 1$ *and is increasing in* A.
$\hat{V}_k(A)$ *is nonincreasing in* k *for fixed* A.

Proof. We only consider $1 < pu + (1-p)d$. Define

$$f(\alpha) = p^{1-\mu}\alpha^{\mu} + (1-p)^{1-\mu}(1-\alpha)^{\mu},\ 0 < \alpha < 1.$$

f takes minimum at $\alpha = p$, $f(p) = 1$ and f is decreasing on $(0, p]$ and increasing on $[p, 1)$.
Given A,

$$(1-\lambda_1) \le A \le (1+\lambda_0).$$

We consider

$$\inf\{f(\alpha)\}.$$

The infimum is taken over α such that there are A^u, A^d satisfying

$$\alpha u A^u + (1-\alpha)dA^d = A,$$

and

$$(1-\lambda_1) \le A^u \le (1+\lambda_0),$$
$$(1-\lambda_1) \le A^d \le (1+\lambda_0).$$

We consider the cases,

(i) $(1-\lambda_1) \le A \le (1-\lambda_1)u$;
(ii) $(1-\lambda_1)u \le A \le (1+\lambda_0)d$;
(iii) $(1+\lambda_0)d \le A \le (1+\lambda_0)$.

Assume (i),

$$\alpha = \frac{A - dA^d}{uA^u - dA^d}.$$

For each $(1-\lambda_1) \le A^u \le (1+\lambda_0)$, the range of α defined above taken over

$$(1-\lambda_1)d \le dA^d \le A$$

is $[0, (A-(1-\lambda_1)d)/(uA^u-(1-\lambda_1)d)$. Take the union of these sets over all

$$(1-\lambda_1) \le A^u \le (1+\lambda_0),$$

we have $[0, (A-(1-\lambda_1)d)/(1-\lambda_1)(u-d)]$. If p is in this interval, then $\hat{V}_{T-1}(A) = 1$. The condition p is in this interval is the same as

$$A \ge (1-\lambda_1)(pu + (1-p)d).$$

Therefore,

$$\hat{V}_{T-1}(A) = 1, (1-\lambda_1)(pu + (1-p)d) \le A \le (1-\lambda_1)u.$$

On the other hand, if

$$(1-\lambda_1) \le A \le (1-\lambda_1)(pu + (1-p)d),$$

the infimum of $f(\alpha)$ on

$$[0, (A-(1-\lambda_1)d)/(1-\lambda_1)(u-d)]$$

is

$$f(\frac{A-(1-\lambda_1)d}{(1-\lambda_1)(u-d)}).$$

Therefore,

$$\hat{V}_{T-1}(A) = f(\frac{A-(1-\lambda_1)d}{(1-\lambda_1)(u-d)})$$

if

$$(1-\lambda_1) \le A \le (1-\lambda_1)(pu + (1-p)d).$$

Assume (ii). We consider $A \le uA^u \le (1+\lambda_0)u$. The range of α is given by $[0, 1]$. Therefore, $\hat{V}_{T-1}(A) = 1$.

Assume (iii). For each $A \le uA^u \le (1+\lambda_0)u$, the range of α of

$$(1-\lambda_1) \le A^d \le (1+\lambda_0)$$

is

$$[\frac{A-(1+\lambda_0)d}{uA^u-(1+\lambda_0)d}, \frac{A-(1-\lambda_1)d}{uA^u-(1-\lambda_1)d}].$$

Take the union of these sets over all A^u gives

$$[\frac{A-(1+\lambda_0)d}{(1+\lambda_0)(u-d)}, 1].$$

We can check p is in this set. Then $\hat{V}_{T-1}(A) = 1$.
We conclude

$$\hat{V}_{T-1}(A) = f(\frac{A-(1-\lambda_1)d}{(1-\lambda_1)(u-d)})$$

if

$$(1-\lambda_1) \le A \le (1-\lambda_1)(pu + (1-p)d),$$

and

$$\hat{V}_{T-1}(A) = 1$$

if

$$(1-\lambda_1)(pu+(1-p)d) \le A \le (1+\lambda_0).$$

$\hat{V}_{T-1}(A)$ is decreasing in A.

We can continue this argument for other $\hat{V}_k(A)$ to prove that $\hat{V}_k(A)=1$ if

$$(1-\lambda_1)(pu+(1-p)d)^{T-k} \le A \le (1+\lambda_0),$$

and for other A, $\hat{V}_k(A) > 1$. To prove the nonincreasing of $\hat{V}_k(A)$ in A needs additional argument. We have the following observation. Let g be nonincreasing. Consider

$$\hat{g}(A) = \inf\{p^{1-\mu}\alpha^\mu g(A^u) + (1-p)^{1-\mu}(1-\alpha)^\mu g(A^d)\},$$

where the "inf" is taken over $0<\alpha<1$ and A^u, A^d satisfying (7.3)′ and (7.4)′. We define

$$\bar{g}(A) = g(A),\ (1-\lambda_1) \le A \le (1+\lambda_0),$$
$$\bar{g}(A) = \infty,\ A < (1-\lambda_1),$$
$$\bar{g}(A) = g((1+\lambda_0)),\ A > (1+\lambda_0).$$

We claim

$$\hat{g}(A) = \inf\{p^{1-\mu}\alpha^\mu \bar{g}(A^u) + (1-p)^{1-\mu}(1-\alpha)^\mu \bar{g}(A^d)\}, \tag{7.5}$$

where the "inf" is taken over $0<\alpha<1$ and A^u, A^d satisfying (7.4)′. First, it is easy to see that the quantity defined by the righthand side of (7.5) is not smaller than $\hat{g}(A)$. To prove the opposite inequality, we observe that for a given $0<\alpha<1$ and A^u, A^d satisfying (7.4)′, if (7.3)′ does not hold, says

$$A^u > (1+\lambda_0).$$

We define $\bar{A}^u = (1+\lambda_0)$ and $\bar{A}^d$ by the relation,

$$\alpha u(1+\lambda_0) + (1-\alpha)d\bar{A}^d = A.$$

Then $\bar{A}^d > A^d$. We see $\bar{A}^u, \bar{A}^d$ satisty (7.3)′ and (7.4)′ and

$$\begin{aligned} &p^{1-\mu}\alpha^\mu \bar{g}(A^u) + (1-p)^{1-\mu}(1-\alpha)^\mu \bar{g}(A^d) \\ &\quad \ge p^{1-\mu}\alpha^\mu g(\bar{A}^u) + (1-p)^{1-\mu}(1-\alpha)^\mu g(\bar{A}^d) \end{aligned}$$

by the property that g is nonincreasing. Using this observation, we can deduce that the quantity defined by the righthand side of (7.5) is not smaller than $\hat{g}(A)$.

Now from (7.5) it is easy to see that $\hat{g}$ is nonincreasing. In fact, let $B=\lambda A>0$ for a $\lambda>1$. Let $0<\alpha<1$ and $A^u, A^d>0$ satisfying

$$\alpha u A^u + (1-\alpha)dA^d = A.$$

We take $B^u = A^u\lambda, B^d = A^d\lambda$. Then

$$\alpha u B^u + (1-\alpha)dB^d = B.$$

We have

$$\begin{aligned} &p^{1-\mu}\alpha^\mu \bar{g}(A^u) + (1-p)^{1-\mu}(1-\alpha)^\mu \bar{g}(A^d) \\ &\quad \ge p^{1-\mu}\alpha^\mu g(B^u) + (1-p)^{1-\mu}(1-\alpha)^\mu g(B^d) \\ &\quad \ge \hat{g}(B). \end{aligned}$$

This is true for any α, A^u, A^d. Therefore, $\hat{g}(A) \geq \hat{g}(B)$.

Finally, we denote $\hat{g}(A) = Hg(A)$. Then H has the property that $g_1(A) \geq g_2(A)$ for all A implies $Hg_1(A) \geq Hg_2(A)$ for all A. Take $g = 1$. Then

$$Hg = \hat{V}_{T-1}.$$

We have proved $\hat{V}_{T-1} \geq 1$. That is,

$$Hg \geq g.$$

We note

$$\hat{V}_k = H\hat{V}_{k+1}.$$

From these, by induction, we can show $\hat{V}_k \geq \hat{V}_{k+1}$. This completes the proof. □

Corollary 7. *Assume* $1 < pu + (1-p)d$ *and*

$$(1-\lambda_1)(pu + (1-p)d)^T \leq (1+\lambda_0).$$

If $x_1 \leq 0$*, then buy-and-hold is an optimal strategy.*

Similarly, assume $1 > pu + (1-p)d$ *and*

$$(1+\lambda_0)(pu + (1-p)d)^T \geq (1-\lambda_1).$$

If $x_1 \geq 0$*, then sell-and-hold is an optimal strategy.*

Proof. Assume $1 < pu + (1-p)d$ and $x_1 \leq 0$. From Theorem 5 and 6,

$$V(x_0, x_1) = \frac{1}{\gamma}(x_0 + x_1 S(1+\lambda_0))^\gamma.$$

Buy-and-hold achives this value and hence is an optimal strategy. Other result can be proved similarly. □

References

[1] Bensaid, B., Lesne, J., Pages, H. and Scheinkman, H. (1992). Derivative asset pricing with transaction costs. *Math. Finance* **2** 63–86.

[2] Boyle, P. and Vorst, T. (1992). Option pricing in discrete time with transaction costs. *J. Finance* **47** 271–293.

[3] Cvitanic, J. and Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs. *Math. Finance* **6** 133–165.

[4] Cvitanic, J., Pham, H. and Touzi, N. (1999). A closed-form solution to the problem of super-replication under transaction costs. *Finance and Stochastics* **3** 35–54.

[5] Chiang, T. S. and Sheu, S. J. (2004). A geometric approach to option pricing with transaction costs in discrete models. Preprint.

[6] Chiang, T. S., Sheu, S. J. and Shih, S. Y. (2005). Dynamics of price systems for multinomial models with transaction costs. Preprint.

[7] Davis, M. H. A., Panas, V. J. and Zariphopoulou, T. (1992). European options pricing with transaction costs. *SIAM J. Control Optim.* **31** 470–498.

[8] Delbaen, F., Kabanov, Y. M. and Valkeila, E. (2002). Hedging under transaction costs in currency markets: a discrete time model. *Math. Finance* **12** 45–61.

[9] Kabanov, Y. M. (1999). Hedging and liquidation under transaction costs in currency markets. *Finance and Stochastics* **3** 237–248.
[10] Kabanov, Y. M. and Last, G. (2002). Hedging under transaction costs in currency markets: A continuous time model. *Math. Finance* **12** 63–70.
[11] Karatzas, I. and Shreve, S. E. (1998). *Methods of Mathematical Finance.* Springer.
[12] Kusuoka, S. (1995). Limiting theorem on option replication with transaction costs. *Ann. Appl. Probab.* **5** 198–221.
[13] Leland, H. E. (1985). Option pricing and replication with transaction costs. *J. Finance* **40** 1283–1301.
[14] Palmer, K. (2004). Replication and super replicating portfolios in the Boyle-Vorst discrete time in option pricing models with transaction costs. Preprint.
[15] Rutowski, M. (1998). Optimality of replication in CRR model with transaction costs. *Applicationes Mathematicae* **25** 29–53.
[16] Settner, L. (1997). Option pricing in CRR model with proportional transaction costs: A cone transformation approach. *Applicationes Mathematicae* **24** 475–514.
[17] Soner, M., Shreve, S. and Cvitanic, J. (1995). There is no nontrivial hedging portfolio for option pricing with transaction costs. *Ann. Appl. Probab.* **5** 327–355.
[18] Touzi, N. (1999). Super-replication under proportional transaction costs: From discrete to continuous-time models. *Math. Meth. Oper. Res.* **50** 297–320.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 272–283

DOI: 10.1214/074921706000001102

A note on the estimation of extreme value distributions using maximum product of spacings

T. S. T. Wong[1] **and W. K. Li**[1]

The University of Hong Kong

Abstract: The maximum product of spacings (MPS) is employed in the estimation of the Generalized Extreme Value Distribution (GEV) and the Generalized Pareto Distribution (GPD). Efficient estimators are obtained by the MPS for all γ. This outperforms the maximum likelihood method which is only valid for $\gamma < 1$. It is then shown that the MPS gives estimators closer to the true parameters compared to the maximum likelihood estimates (MLE) in a simulation study. In cases where sample sizes are small, the MPS performs stably while the MLE does not. The performance of MPS estimators is also more stable than those of the probability-weighted moment (PWM) estimators. Finally, as a by-product of the MPS, a goodness of fit statistic, Moran's statistic, is available for the extreme value distributions. Empirical significance levels of Moran's statistic calculated are found to be satisfactory with the desired level.

1. Introduction

The GEV and the GPD (Pickands, [13]) distributions are widely-adopted in extreme value analysis. As is well known the maximum likelihood estimates (MLE) may fail to converge owing to the existence of an unbounded likelihood function. In some cases, MLE can be obtained but converges at a slower rate when compared to that of the classical MLE under regular conditions.

Recent studies (e.g. Juarez & Schucany, [9]) show that maximum likelihood estimation and other common estimation techniques lack robustness. In addition, the influence curve of the MLE is shown unstable when the sample size is small. Although new methods (Juarez & Schucany, [9]; Peng and Welsh, [12]; Dupuis, [6]) were proposed, arbitrary parameters are sometimes involved, resulting in more intensive computation which is in general undesirable. There have been studies in overcoming the difficulties of the MLE in extreme value analysis but none has considered the MPS. Furthermore, a goodness-of-fit test on the fitted GEV or GPD is rarely considered.

In this study, the MPS method will first be considered for the purpose of finding estimators which may not be obtained by the maximum likelihood method. As a by product, the Moran's statistic, a function of product of spacings, can be treated as a test statistics for model checking. This is one of the nice outcomes of MPS which Cheng and Stephens [4] demonstrated but is overlooked by the extreme value analysis literature.

In Section 2, we discuss some problems of the MLE. In Section 3, we formulate the MLE, the MPS and the Moran statistics. In Section 4, results of simulation studies

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, e-mail: h0127272@hkusua.hku.hk

Keywords and phrases: generalized extreme value distribution, generalized Pareto distribution, maximum product of spacings, maximum likelihood, Moran's statistic.

are presented to evaluate the performance of the method proposed. In Section 5, we provide some real examples in which the MPS is more convincing. A brief discussion is presented in Section 6.

2. Problems of the MLE

The problems of the MLE in model fitting were discussed by Weiss and Wolfowitz [15]. Related discussions in connection to the Weibull and the Gamma distributions can be found in [2, 3, 5, 14]. Smith [14] found densities in the form

$$f(x;\theta,\phi) = (x-\theta)^{\alpha-1} g(x-\theta;\phi), \quad \theta < x < \infty \tag{2.1}$$

where θ and ϕ are unknown parameters and g converges to a constant as $x \downarrow \theta$. As is well-known for $\alpha > 2$, the MLE is as efficient as in regular cases. For $\alpha = 2$, the estimated parameters are still asymptotically normal, but the convergence rate is $(n \log n)^{\frac{1}{2}}$ which is larger than the classical rate of $n^{\frac{1}{2}}$. For $1 < \alpha < 2$, the MLE exists but the asymptotic efficiency problem is not solved. And the order of convergence could be as high as $O(n^{\frac{1}{\alpha}})$. For $\alpha < 1$, MLE does not exist. Both the GEV and the GPD encounter the above difficulties as both can be reparameterised into the form (2.1).

As an alternative to the MLE, the MPS was established by Cheng and Amin [2]. With the MPS, not only can problems with non-regular condition be better solved, but models originally estimable under the MLE framework can also be better estimated by the MPS using a much simpler algorithm. Cheng & Amin [2] showed that the MPS estimators are asymptotically normal even for $0 < \alpha < 1$. This overcomes to a certain extent the weakness existing in the MLE. Hence, the MPS may be one of the most robust estimation techniques and yet the least computational expensive in extreme value analysis. The present paper employs the MPS in the estimation of the GEV and the GPD. On the other hand, many previous studies (Hosking, [7]; Marohn, [10]) concentrated on testing the shape parameter. Goodness-of-fit test on the model as a whole has been very few. In this study, the Moran's statistic (Cheng and Stephens, [4]; Moran, [11]) arising naturally as a by product of the MPS estimator was utilized to check the adequacy of the overall model.

3. Formulations of the MLE, the MPS and the Moran's statistic

3.1. The MLE and the MPS

The c.d.f of the GEV and the GPD are respectively

$$H(x;\gamma,\mu,\sigma) = \exp\left[-\left(1-\gamma\frac{x-\mu}{\sigma}\right)^{\frac{1}{\gamma}}\right], \quad 1-\gamma\frac{x-\mu}{\sigma} > 0;$$

and

$$G(x;\gamma,\sigma) = 1-\left(1-\gamma\frac{x}{\sigma}\right)^{\frac{1}{\gamma}}, \quad 1-\gamma\frac{x}{\sigma} > 0.$$

where

$$\gamma \neq 0, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

Let $h(x)$ and $g(x)$ be the corresponding densities,

$$h(x) = \frac{1}{\sigma}\left(1-\gamma\frac{x-\mu}{\sigma}\right)^{\frac{1}{\gamma}-1}\exp\left[-\left(1-\gamma\frac{x-\mu}{\sigma}\right)^{\frac{1}{\gamma}}\right] ;$$

and

$$g(x) = \frac{1}{\sigma}\left(1-\gamma\frac{x}{\sigma}\right)^{\frac{1}{\gamma}-1} .$$

The log-likelihood functions per observation are respectively

$$L_{\mathrm{GEV}}(\gamma,\mu,\sigma) = -\log\sigma + \left(\frac{1}{\gamma}-1\right)\log\left(1-\gamma\frac{x-\mu}{\sigma}\right) - \left(1-\gamma\frac{x-\mu}{\sigma}\right)^{\frac{1}{\gamma}} ;$$

and

$$L_{\mathrm{GPD}}(\gamma,\mu,\sigma) = -\log\sigma + \left(\frac{1}{\gamma}-1\right)\log\left(1-\gamma\frac{x}{\sigma}\right) .$$

Applying the same argument stated in [14], as $x \downarrow \mu+\frac{\sigma}{\gamma}$, the information matrix of $L_{\mathrm{GEV}}(\gamma,\mu,\sigma)$ is infinite for $\gamma > \frac{1}{2}$. The same difficulty arises in the GPD as $x \downarrow \frac{\sigma}{\gamma}$. In this case, the underlying distribution is J-shaped where maximum likelihood is bound to fail. Worse still, MLEs (Denoted by $\hat{\boldsymbol{\Theta}}_{\mathrm{GEV}} = \left(\hat\gamma,\hat\mu,\hat\sigma\right)^T$ and $\hat{\boldsymbol{\Theta}}_{\mathrm{GPD}} = \left(\hat\gamma,\hat\sigma\right)^T$ respectively for the GEV and the GPD) may not exist when $\gamma > 1$. Let $x_1 < x_2 < \cdots < x_n$ be an ordered sample of size n and define spacings $D_i(\boldsymbol{\theta})$ by

$$\begin{aligned}
&\mathrm{GEV}: D_i(\boldsymbol{\theta}) = H(x_i,\gamma,\mu,\sigma) - H(x_{i-1};\gamma,\mu,\sigma) , \quad (i = 1,2,\ldots,n+1) ;\\
&\mathrm{GPD}: D_i(\boldsymbol{\theta}) = G(x_i,\gamma,\sigma) - G(x_{i-1};\gamma,\sigma) , \qquad\quad (i = 1,2,\ldots,n+1) ;
\end{aligned}$$

where $H(x_0;\gamma,\mu,\sigma) \equiv G(x_0;\gamma,\sigma) \equiv 0$ and $H(x_{n+1};\gamma,\mu,\sigma) \equiv G(x_{n+1};\gamma,\sigma) \equiv 1$.

MPS estimators (Denoted by $\breve{\boldsymbol{\Theta}}_{\mathrm{GEV}} = (\breve\gamma,\breve\mu,\breve\sigma)^T$ and $\breve{\boldsymbol{\Theta}}_{\mathrm{GPD}} = (\breve\gamma,\breve\sigma)^T$ respectively for the GEV and the GPD) are found by minimizing

$$M(\boldsymbol{\theta}) = -\sum_{i=1}^{n+1}\log D_i(\boldsymbol{\theta}).$$

By taking the cumulative density in the estimation, the objective function $M(\boldsymbol{\theta})$ does not collapse for $\gamma < 1$ as $x \downarrow \mu+\frac{\sigma}{\gamma}$ for the GEV or as $x \downarrow \frac{\sigma}{\gamma}$ for the GPD. The MLE, however, does not have such an advantage. There is in probability a solution $\hat{\boldsymbol{\Theta}}$ that is asymptotically normal only for $\gamma < \frac{1}{2}$. The strength of MPS over MLE is demonstrated by the following two theorems.

Theorem 3.1. *Let $\boldsymbol{\Theta}_{0\mathrm{GEV}} = (\gamma_0,\mu_0,\sigma_0)^T$ and $\boldsymbol{\Theta}_{0\mathrm{GPD}} = (\gamma_0,\sigma_0)^T$ be the true parameters of the GEV and the GPD respectively. Under regularity conditions (See for example:* [14]*)*

(i) *For $\gamma < \frac{1}{2}$, $n^{\frac{1}{2}}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D} N\Big(\mathbf{0}, -E\Big(\frac{\partial^2 L}{\partial\boldsymbol{\Theta}^2}\Big)^{-1}\Big)$;*

(ii) *For $\gamma = \frac{1}{2}$, $\Big(\hat\mu + \frac{\hat\sigma}{\hat\gamma}\Big) - (\mu_0 + \frac{\sigma_0}{\gamma_0}) \xrightarrow{D} O_p[(n\log n)^{-\frac{1}{2}}]$, and $n^{\frac{1}{2}}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D}$ $N(\mathbf{0}, -E(\frac{\partial^2 L}{\partial\boldsymbol{\Theta}^2})^{-1})$, where $\boldsymbol{\Theta} = (\gamma,\sigma)^T$;*

(iii) *For $\frac{1}{2} < \gamma < 1$, $\Big(\hat\mu + \frac{\hat\sigma}{\hat\gamma}\Big) - \Big(\mu_0 + \frac{\sigma_0}{\gamma_0}\Big) \xrightarrow{D} O_p(n^{-\gamma})$, and $n^{\frac{1}{2}}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D}$ $N\Big(\mathbf{0}, -E\Big(\frac{\partial^2 L}{\partial\boldsymbol{\Theta}^2}\Big)^{-1}\Big)$, where $\boldsymbol{\Theta}$ is as in* (ii).

(iv) *For* $\boldsymbol{\gamma} \geq \mathbf{1}$, *the MLE does not exist.*

Theorem 3.2. *Under the same conditions as in Theorem 3.1*

(i) *For* $\gamma < \frac{1}{2}$, $n^{\frac{1}{2}}(\breve{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D} N(\mathbf{0}, -E(\frac{\partial^2 L}{\partial \boldsymbol{\Theta}^2})^{-1})$;

(ii) *For* $\gamma = \frac{1}{2}$, $(\breve{\mu} + \frac{\breve{\sigma}}{\breve{\gamma}}) - (\mu_0 + \frac{\sigma_0}{\gamma_0}) \xrightarrow{D} O_p[(n \log n)^{-\frac{1}{2}}]$, *and* $n^{\frac{1}{2}}(\breve{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D} N(\mathbf{0}, -E(\frac{\partial^2 L}{\partial \boldsymbol{\Theta}^2})^{-1})$, *where* $\boldsymbol{\Theta} = (\gamma, \sigma)^T$;

(iii) *For* $\gamma > \frac{1}{2}$, $(\breve{\mu} + \frac{\breve{\sigma}}{\breve{\gamma}}) - (\mu_0 + \frac{\sigma_0}{\gamma_0}) \xrightarrow{D} O_p(n^{-\gamma})$, *and* $n^{\frac{1}{2}}(\breve{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{D} N(\mathbf{0}, -E(\frac{\partial^2 L}{\partial \boldsymbol{\Theta}^2})^{-1})$, *where* $\boldsymbol{\Theta}$ *is as in* (ii).

Proofs of Theorems 3.1 and 3.2 follow the arguments in [14] and [2] respectively by checking the conditions therein.

It is obvious that efficient estimators can still be obtained by the MPS for $\gamma > \frac{1}{2}$ but not the MLE. From (iii) above, it is clear that the MPS still works while the MLE fails for $\gamma \geq 1$. It seems that it is a fact overlooked by researchers working in the extreme value literature.

3.2. Moran's statistic

In the MPS estimation, $M(\boldsymbol{\theta})$ is called the Moran's statistic which can be used as a test for a goodness-of-fit test. Cheng and Stephens [4] showed that under the null hypothesis, $M(\boldsymbol{\theta})$, being independent of the unknown parameters, has a normal distribution and a chi-square approximation exists for small samples with mean and variance approximated respectively by

$$\mu_M \approx (n+1)\log(n+1) - \frac{1}{2} - \frac{1}{12(n+1)} \ ,$$

and

$$\sigma_M^2 \approx (n+1)\left(\frac{\pi^2}{6} - 1\right) - \frac{1}{2} - \frac{1}{6(n+1)} \ .$$

Define

$$C_1 = \mu_M - \left(\frac{1}{2}n\right)^{\frac{1}{2}}\sigma_M \ , \quad C_2 = (2n)^{-\frac{1}{2}}\sigma_M \ .$$

The test statistic is

$$T(\breve{\boldsymbol{\theta}}) = \frac{M(\breve{\boldsymbol{\theta}}) + \frac{1}{2}k - C_1}{C_2}$$

which follows approximately a chi-square distribution of n degrees of freedom under the null hypothesis. Monte Carlo simulation of the Weibull, the Gamma and the Normal distributions in [4] showed the accuracy of the test based on $T(\breve{\boldsymbol{\theta}})$. In the next section, we provide further evidence supporting the use of MPS for fitting the extreme value distributions.

4. Simulation study

A set of simulations was performed to evaluate the advantage of the MPS over the MLE of the GEV and the GPD based on selected parameters for different sample

sizes $n = (10, 20, 50)$. Empirical significance levels of Moran's statistic were then considered using $\chi^2_{n,\alpha}$ as the benchmark critical value. Finally, data were generated from an exponential distribution and the cluster maxima of every 30 observations were fitted to the GEV.

The subroutine DNCONF in the IMSL library was used to minimize a function. The data analysed in the paper and the Fortran90 programs used in the computation are available upon request.

We have done extensive simulations to assess the performance of MPS estimators. Only four simulation results in each combination of γ and n are reported. The location and scale parameter, $\mu = 1$ and $\sigma = 1$, were used throughout. On the basis of the results from asymptotic normality of the MPS that were presented in Section 3, we chose a combination of $\gamma = (-0.2, 0.2, 1, 1.2)$ to compare the estimation performance between the maximum likelihood method and the MPS where the last two cases should break down for the MLE. 10000 simulations of sample sizes $n = (10, 20, 50)$ were performed. Data were generated from the same random seed and estimations were performed under the same algorithm. Define the mean absolute error for the MLE and the MPS respectively by

$$\frac{1}{l}\left|\hat{\boldsymbol{\Theta}}_l^T - \boldsymbol{\Theta}_0 \mathbf{1}^T\right|\mathbf{1} \quad \text{and} \quad \frac{1}{l}\left|\breve{\boldsymbol{\Theta}}_l^T - \boldsymbol{\Theta}_0 \mathbf{1}^T\right|\mathbf{1} \ .$$

where $\hat{\boldsymbol{\Theta}}_l$ and $\breve{\boldsymbol{\Theta}}_l$ are $l \times 1$ vectors of the MLE and MPS estimators respectively, $|\boldsymbol{Y}|$ means the element-wise absolute value of $\boldsymbol{Y}$, p is the number of estimated parameters and $l = 10000$ is the number of replications. The mean absolute error measures the average deviation of estimators from the true parameters and hence is a measure of robustness. A small mean absolute error is expected.

As suggested by a referee, the MPS was also compared to the method of probability-weighted moment (PWM)(Hosking et al., [8]) for the GEV model. We followed Hosking's approach in his Table 3 and estimated the tail parameter by Newton-Raphson's Method. Tables 1 and 2 display the medians of the parameters in 10000 estimations together with the mean absolute error in bracket. Both the MPS estimates and the MLEs are in line with the true parameters but MPS tends to give a closer result for the GEV. It can also be seen that the MPS gives much more stable estimates than the MLE in general. For $\gamma = -0.2$ and $\gamma = 0.2$, the PWM performed well with slightly smaller mean absolute errors than the MPS. However, for $\gamma = 1$ and $\gamma = 1.2$, the bias of the PWM is rather severe. Note that some of the mean absolute errors for the MLE are unacceptably large due to serious outliers of estimated parameters. Non-regularity of the likelihood function caused occasional non-convergence. The frequency of such problems is reported in Tables 3 and 4. Failures of convergence were detected when the magnitudes of any estimator in an entry exceeds 100. The failure rates of MLE are relatively higher than those of MPS. Some estimated parameters of the MLE went up to as high as 500000. This explains the extremely large mean absolute errors of the MLE. Although there were failures in MPS, the maximum values were less than 1000, comparably less severe than the MLE. The PWM has zero failure rates but as mentioned above, it has a severe bias when $\gamma \geq 1$.It is noticed that the MLEs have smaller mean absolute error only in cases where sample size is large. However, the MPS estimators have virtually no fall off in its performance across sample sizes. These are in agreement with the theoretical results in Theorems 3.1 and 3.2. Overall, the MPS seems to be the most stable in its performance.

The Moran's statistic, $M(\boldsymbol{\theta})$, has a chi-square distribution with n degrees of freedom. Monte Carlo simulations with 10000 observations per entry, each entry with

TABLE 1

Simulation results of MPS estimates, MLEs and PWM estimates on the GEV. Shown are medians of estimated parameters from 10000 *simulations of sample sizes* $n = (10, 20, 50)$. *Numbers in the bracket are mean absolute errors of estimates*

n	True parameters			MPS estimates			MLEs			PWM estimates		
	γ_0	μ_0	σ_0	$\breve{\gamma}$	$\breve{\mu}$	$\breve{\sigma}$	$\hat{\gamma}$	$\hat{\mu}$	$\hat{\sigma}$	γ	μ	σ
10	−0.2	1	1	−0.28	0.96	1.06	−0.22	1.09	0.98	−0.15	1.02	0.93
				(0.38)	(0.30)	(0.30)	(435)	(175)	(295)	(0.18)	(0.29)	(0.25)
	0.2	1	1	0.20	0.97	1.09	0.43	0.98	0.84	0.10	0.97	0.88
				(0.33)	(0.30)	(0.27)	(117)	(69)	(104)	(0.18)	(0.28)	(0.20)
	1	1	1	1.17	0.98	1.11	1.20	0.78	0.78	0.59	0.89	0.88
				(0.53)	(0.33)	(0.45)	(54)	(90)	(50)	(0.42)	(0.31)	(0.27)
	1.2	1	1	1.40	0.99	1.13	1.36	0.77	0.80	0.70	0.87	0.90
				(0.90)	(0.35)	(0.49)	(32)	(50)	(0.26)	(0.50)	(0.33)	(0.31)
20	−0.2	1	1	−0.25	0.97	1.06	−0.26	1.06	1.02	−0.17	1.01	0.96
				(0.20)	(0.21)	(0.20)	(0.69)	(0.98)	(0.47)	(0.14)	(0.21)	(0.17)
	0.2	1	1	0.20	0.98	1.07	0.25	1.09	1.00	0.14	0.98	0.94
				(0.17)	(0.20)	(0.17)	(46)	(1.01)	(50)	(0.13)	(0.20)	(0.14)
	1	1	1	1.10	0.99	1.09	1.18	0.80	0.80	0.77	0.94	0.95
				(0.36)	(0.24)	(0.27)	(85)	(34)	(95)	(0.27)	(0.21)	(0.20)
	1.2	1	1	1.33	0.99	1.10	1.35	0.79	0.81	0.92	0.93	0.96
				(0.55)	(0.26)	(0.31)	(13)	(50)	(0.33)	(0.32)	(0.22)	(0.22)
50	−0.2	1	1	−0.22	0.99	1.04	−0.25	1.02	1.0	−0.18	1.01	0.98
				(0.11)	(0.13)	(0.11)	(0.12)	(0.13)	(0.11)	(0.10)	(0.13)	(0.11)
	0.2	1	1	0.20	0.99	1.04	0.20	1.04	1.01	0.18	0.99	0.97
				(0.09)	(0.13)	(0.10)	(0.09)	(0.13)	(0.09)	(0.08)	(0.13)	(0.09)
	1	1	1	1.05	0.99	1.05	1.11	0.89	0.88	0.90	0.97	0.98
				(0.29)	(0.16)	(0.16)	(50)	(50)	(0.22)	(0.16)	(0.13)	(0.12)
	1.2	1	1	1.27	0.99	1.06	1.29	0.87	0.89	1.08	0.97	0.99
				(0.14)	(0.12)	(0.17)	(0.15)	(0.15)	(0.15)	(0.19)	(0.13)	(0.14)

TABLE 2

Simulation results of MPS estimates and MLEs on the GPD. Shown are medians of estimated parameters from 10000 *simulations of sample sizes* $n = (10, 20, 50)$. *Numbers in the bracket are mean absolute errors of estimates*

n	True parameters		MPS estimates		MLEs	
	γ_0	μ_0	$\breve{\gamma}$	$\breve{\mu}$	$\hat{\gamma}$	$\hat{\mu}$
10	−0.2	1	−0.10(0.48)	0.97(0.44)	−0.52(170)	0.80(247)
	0.2	1	0.41(0.55)	0.92(0.48)	−0.18(7547)	0.99(12347)
	1	1	1.33(0.75)	0.86(0.57)	0.72(5304)	1.20(17860)
	1.2	1	1.58(0.81)	0.83(0.59)	0.92(3890)	1.17(14582)
20	−0.2	1	−0.13(0.26)	0.98(0.28)	−0.45(0.27)	0.97(0.28)
	0.2	1	0.33(0.31)	0.95(0.31)	−0.03(506)	1.06(950)
	1	1	1.21(0.46)	0.91(0.37)	0.87(70)	1.08(300)
	1.2	1	1.43(0.50)	0.90(0.39)	1.06(10)	1.07(50)
50	−0.2	1	−0.15(0.13)	0.98(0.16)	−0.38(50)	0.99(24)
	0.2	1	0.27(0.18)	0.96(0.21)	0.07(50)	1.02(50)
	1	1	1.11(0.26)	0.95(0.23)	0.95(0.23)	1.02(0.24)
	1.2	1	1.32(0.28)	0.94(0.24)	1.14(0.25)	1.03(0.25)

sample size $n = (10, 20, 50)$ were conducted to compute the empirical significant levels. Again the null distributions were the models under consideration in Tables 1 and 2. It can be seen from Tables 5 and 6 that the empirical sizes for both the GEV and the GPD are very conservative at small sample sizes $n = 10$. Improvement was seen at $n = 20$. Though slightly conservative, it is acceptable in some applications. But the results at $n = 50$ are very good even with $\gamma = 1$ and $\gamma = 1.2$.

We have also evaluated the empirical significance level of models having different

TABLE 3
Failure rate of MPS estimation, maximum likelihood estimation and PWM estimation for the GEV Distribution. Tabulated values are the number of outliers per 100 *simulated samples*

n	MPS estimation				maximum likelihood estimation				PWM estimation			
	γ_0				γ_0				γ_0			
	-0.2	0.2	1	1.2	-0.2	0.2	1	1.2	-0.2	0.2	1	1.2
10	0.00	0.00	0.03	0.05	2.00	0.77	0.11	0.02	0.00	0.00	0.00	0.00
20	0.00	0.00	0.06	0.11	0.06	0.03	0.04	0.03	0.00	0.00	0.00	0.00
50	0.00	0.00	0.05	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00

TABLE 4
Failure rate of MPS estimation and maximum likelihood estimation for the GPD distribution. Tabulated values are the number of outliers per 100 *simulated samples*

n	MPS estimation				maximum likelihood estimation			
	γ_0				γ_0			
	-0.2	0.2	1	1.2	-0.2	0.2	1	1.2
10	0.00	0.00	0.00	0.00	0.05	2.51	3.61	2.94
20	0.00	0.00	0.00	0.00	0.00	0.19	0.06	0.01
50	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00

TABLE 5
Empirical sizes of Moran test statistics on the GEV from 10000 *simulations of sample sizes* $n = (10, 20, 50)$

n	GEV Models			Empirical sizes		
	γ_0	μ_0	σ_0	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
10	-0.2	1	1	0.0618	0.0270	0.0034
	0.2	1	1	0.0580	0.0257	0.0037
	1	1	1	0.0592	0.0264	0.0062
	1.2	1	1	0.0619	0.0318	0.0115
20	-0.2	1	1	0.0759	0.0337	0.0053
	0.2	1	1	0.0783	0.0373	0.0081
	1	1	1	0.0785	0.0360	0.0086
	1.2	1	1	0.0814	0.0377	0.0089
50	-0.2	1	1	0.0848	0.0408	0.0077
	0.2	1	1	0.0906	0.0414	0.0074
	1	1	1	0.0890	0.0419	0.0101
	1.2	1	1	0.1000	0.0509	0.0143

TABLE 6
Empirical sizes of Moran test statistics on the GPD from 10000 *simulations of sample sizes* $n = (10, 20, 50)$

n	GPD Models		Empirical sizes		
	γ_0	σ_0	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
10	-0.20	1	0.0745	0.0326	0.0044
	0.20	1	0.0699	0.0326	0.0044
	1.00	1	0.0718	0.0326	0.0044
	1.20	1	0.0734	0.0326	0.0044
20	-0.20	1	0.0855	0.0394	0.0070
	0.20	1	0.0827	0.0374	0.0066
	1.00	1	0.0794	0.0375	0.0078
	1.20	1	0.0806	0.0377	0.0071
50	-0.20	1	0.0972	0.0474	0.0093
	0.20	1	0.0932	0.0452	0.0081
	1.00	1	0.0960	0.0471	0.0091
	1.20	1	0.0940	0.0477	0.0087

TABLE 7

GEV parameter estimation by MPS in 10000 *simulations of sample sizes* $n = 10, 20, 50$. *Data are generated from exponential distributions with* $\lambda = 0.1, 0.5, 1.0, 5.0$. *Figures shown are* 25%, 50% *and* 75% *quantiles of the estimated parameter.*

		Parameter quantile estimates of GEV								
n	λ	25%$\breve{\gamma}$	50%$\breve{\gamma}$	75%$\breve{\gamma}$	25%$\breve{\mu}$	50%$\breve{\mu}$	75%$\breve{\mu}$	25%$\breve{\sigma}$	50%$\breve{\sigma}$	75%$\breve{\sigma}$
10	0.1	−0.33	−0.07	0.22	31.51	33.82	36.27	8.33	10.56	12.88
	0.5	−0.33	−0.07	0.21	6.3	6.76	7.25	1.67	2.11	2.58
	1.0	−0.33	−0.07	0.22	3.15	3.38	3.63	0.83	1.06	1.29
	5.0	−0.32	−0.04	0.27	0.62	0.68	0.73	0.17	0.22	0.28
20	0.1	−0.18	−0.05	0.11	32.28	33.93	35.65	8.97	10.39	11.79
	0.5	−0.19	−0.05	0.11	6.46	6.79	7.13	1.79	2.08	2.36
	1.0	−0.19	−0.05	0.11	3.23	3.39	3.57	0.9	1.04	1.18
	5.0	−0.19	−0.05	0.11	0.65	0.68	0.71	0.18	0.21	0.24
50	0.1	−0.11	−0.03	0.05	33.01	34.06	35.14	9.33	10.13	10.99
	0.5	−0.11	−0.03	0.05	6.6	6.81	7.03	1.87	2.03	2.2
	1.0	−0.11	−0.03	0.05	3.3	3.41	3.51	0.93	1.01	1.1
	5.0	−0.11	−0.03	0.05	0.66	0.68	0.70	0.19	0.2	0.22

scale and location parameters. The results did not differ significantly and thus were not reported here. It seems that the performance of Moran's statistic was affected by the sample size rather than the underlying models.

In application, it is common to take cluster maxima in the model fitting of the GEV. Having shown that the MPS gives stable estimations on data generated from known models, in the following, fitting the maximum observations in clusters of size 30 was performed. This experiment mimics the situation that the original data are daily observations with GEV fitted to the monthly maxima. The aim of this experiment is to evaluate the stability in the estimation of cluster maxima.

Data $x_{n,m}$ were simulated from the exponential distribution

$$F(x) = 1 - e^{-\lambda x} \quad x > 0$$

with $\lambda = 0.1, 0.5, 1.0, 5.0$ where n was the sample size of maxima and $m = 30$ the size of a cluster. From each cluster, the maximum, $\max(x_{n,1}, \ldots, x_{n,30})$, was taken and the GEV distributions was fitted to the data by MPS method.

Table 7 shows the estimated parameter quantiles. In the GEV fitting, the tail estimates fall in a narrow range in the four cases $\lambda = 0.1, 0.5, 1.0, 5.0$. Note that the medians for $\breve{\sigma}$ are proportional to the value of λ^{-1}. This feature remains stable across all sample sizes. A similar pattern is also observed for the medians of $\breve{\mu}$. This again shows that estimation using MPS is stable and reliable.

5. Real examples

Some real data sets were studied in the literature (Castillo et al., [1]) using the maximum likelihood method. To illustrate the advantages of the MPS approach, in this paper, four examples were studied, namely, the age data, the wave data, the wind data and the flood data. The above four data sets are obtainable in Castillo et al. [1]. The first example is the oldest age of men at death in Sweden. The annual oldest ages at death in Sweden from 1905 to 1958 were recorded. The age data may be used to predict oldest ages at death in the future. The wave data set contains the yearly maximum heights, in feet. The data could be used in the design of a breakwater. Then, in the wind data, the yearly maximum wind speed in miles per hour is considered. A wind speed design for structural building purposes could be determined from this data set. The last example is the flood data which

TABLE 8
Estimated GPD parameters by MPS in four examples

Data	Threshold	$\breve{\gamma}$	$\breve{\sigma}$	$M(\breve{\theta})$
Age	104.01	1.06	2.79	43.01
Wave	17.36	0.01	7.00	45.81
Flood	45.04	−0.03	9.62	60.53
Wind	36.82	−0.88	5.31	46.79

TABLE 9
Estimated GPD parameters by maximum likelihood method in four examples

Data	Threshold	$\hat{\gamma}$	$\hat{\sigma}$	Log-likelihood
Age	104.01	1.38	3.45	−9.23
Wave	17.36	0.27	7.98	−39.33
Flood	45.04	0.20	10.87	−57.34
Wind	36.82	−0.48	6.52	−47.01

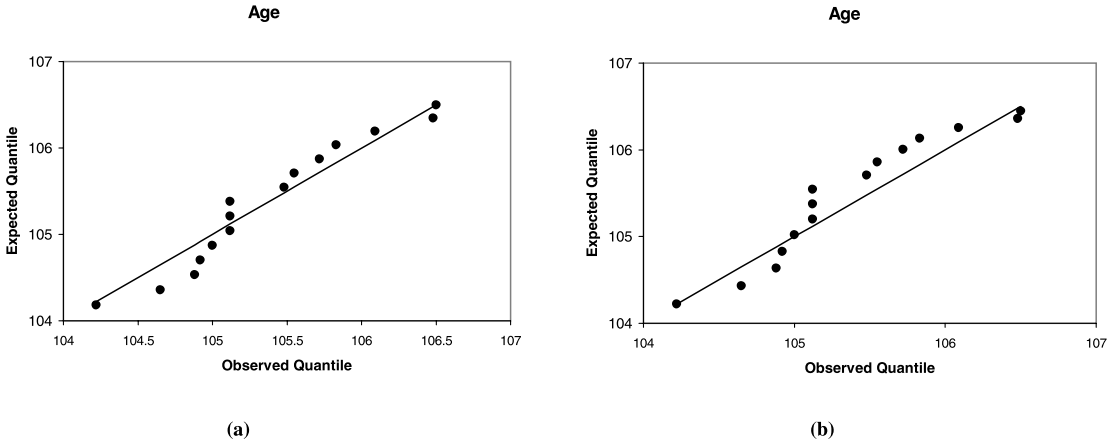


FIG 1. *Quantile plots of the age data fitted with the GPD using the MPS* (a) *and the MLE* (b).

consists of the yearly maximum flow discharge, in cubic meters. The data may help in designing a flood protection device.

In this section, we focus on the GPD with the maximum likelihood method and the MPS method. The GPD was fitted to the excess over a threshold. The thresholds were taken from [1]. Fitted parameters are shown in Tables 8 and 9. Note that $\breve{\gamma}$ and $\hat{\gamma}$ are greater than 1 for the age data. They are less than 1 for the wave, flood and wind data sets.

5.1. The GPD model for age data

Recall from Theorem 3.1 that the MLE does not exist for $\gamma > 1$. When the GPD is fitted to the age data, maximization of the GPD log-likelihood leads to the estimate $(\hat{\gamma}, \hat{\sigma}) = (1.38, 3.45)$ for which the log-likelihood is -9.23. The corresponding values using MPS are $(\breve{\gamma}, \breve{\sigma}) = (1.06, 2.79)$ and $M(\breve{\boldsymbol{\theta}}) = 43.01$. Fig. 1 shows the quantile plot of the two models fitted by the MPS and the maximum likelihood method respectively. In each plot, the expected quantile is calculated by

$$\text{GPD} : Q_{\text{GPD}} = \left[1 - \left(1 - \frac{i}{n+1}\right)^{\gamma}\right] \frac{\sigma}{\gamma} \quad (i = 1, 2, \ldots, n) \ .$$

where $\boldsymbol{\Theta}_{\text{GPD}} = (\gamma, \sigma)^T$ are estimated parameters either by the MPS or by the maximum likelihood method.

The MPS seems to perform better than the MLE. Empirical upper quantiles in the MPS are closer to that of a straight line. This suggests that the MPS is a better method in this case.

5.2. GPD model for the wave data, flood data and wind data

The GPD was also considered for the wave data, the flood data and the wind data. Thresholds for the GPDs were taken as in Castillo et al. [1]. The quantile pots for the MPS are reported in Fig. 2(a), 2(c) and 2(e) and those for the MLE are

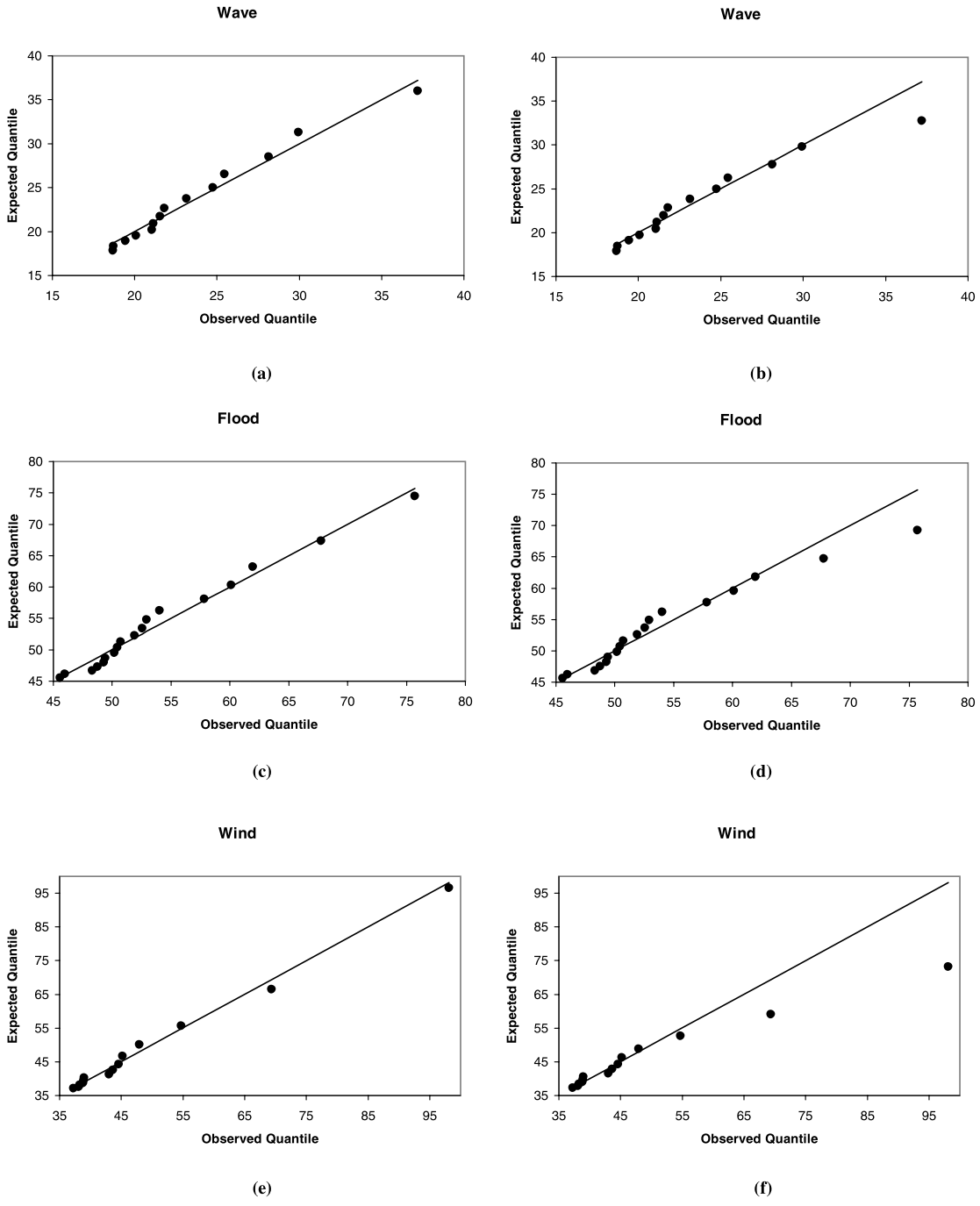


FIG 2. *Quantile plots of the wave data (*(a) *and* (b)*), the flood data (*(c) *and* (d)*) and the wind data (*(e) *and* (f)*) fitted with the GPD. The expected quantiles of* (a), (c) *and* (e) *were based on the MPS. The expected quantiles of* (b), (d) *and* (f) *were based on the MLE.*

reported in Fig. 2(b), 2(d) and 2(f). With reference to Fig. 2(a), 2(c) and 2(e), it can be seen that empirical quantiles based on the MPS keep close to the fitted model's. However, in Fig. 2(b), 2(d) and 2(f), plots of the upper quantiles based on the MLE seem to deviate more from a straight line. This suggests that the MPS gives a better fit to the data.

6. Conclusion and discussion

In extreme value analysis, one technical problem is the lack of data owing to the fact that only extreme observations are used for model fitting. Subject to this constraint, a method that is able to give stable estimates is highly desirable. Juarez and Schucany [9] have demonstrated the instability of the influence curve of the MLE at small sample sizes. This is in agreement with the presented simulation results. In contrast, the MPS works satisfactorily. Not only does the MPS yield closer estimates from data generated from a known parameter set, it also keeps performing stably for data maxima taken from clusters. It also works well under $\gamma \geq 1$ whereas the MLE does not. In addition to MPS's simple formulation and execution, its by-product, the Moran's statistic, is shown to perform well in checking the goodness of fit. The MPS could potentially be one of the best methods in fitting extreme value distributions. On the other hand, it has been shown in [2] that the MPS is a function of sufficient statistics. Extension to multivariate problems using MPS is also going to be explored.

Acknowledgements

W. K. Li would like to thank the organiser of the C. Z. Wei Memorial Conference for the invitation to participate in the conference. Partial support by the Area of Excellence Scheme under the University Grants Committee of the Hong Kong Special Administration Region, China (Project AoE/P-04/2004) is also gratefully acknowledged. The authors thank a referee for comments that led to improvement of the paper.

References

[1] Castillo, E., Hadi, A. S., Balakrishnan, N. and Sarabia, J. M. (2005). *Extreme Value and Related Models with Applications in Engineering and Science*. John Wiley & Sons, Inc., New Jersey.

[2] Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted region. *Journal of Royal Statistics Society B* **45** (3) 394–403.

[3] Cheng, R. C. H. and Iles, T. C. (1987). Corrected maximum likelihood in non-regular problems. *Journal of Royal Statistical Society B* **49** (1) 95–101.

[4] Cheng, R. C. H. and Stephens, M. A. (1989). A goodness-of-fit test using Moran's statistic with estimated parameters. *Biometrika* **76** (2) 385–392.

[5] Cheng, R. C. H. and Traylor, L. (1995). Non-regular maximum likelihood problems. *Journal of Royal Statistics Society B* **57** (1) 3–44.

[6] Dupuis, D. J. (1999). Exceedances over high thresholds: A guide to threshold selection. *Extremes* **1** (3) 251–261.

[7] Hosking, J. R. M. (1984). Testing whether the shape parameter is zero in the generalized extreme value distribution. *Biometrika* **71** (2) 367–374.

[8] Hosking, J. R. M., Wallis, J. R. and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moment. *Technometrics* **27** (3) 251–261.
[9] Juarez, S. F. and Schucany, W. R. (2004). Robust and efficient estimation for the generalized pareto distribution. *Extremes* **7** 237–251.
[10] Marohn, F. (2000). Testing extreme value models. *Extremes* **3** (4) 363–384.
[11] Moran, P. (1951). The random division of an interval – Part II. *Journal of Royal Statistics Society B* **13** 147–150.
[12] Peng, L. and Welsh, A. (2002). Robust estimation of the generalized pareto distribution. *Extremes* **4** (1) 53–65.
[13] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annuals of Statistics* **3** 119–131.
[14] Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72** (1) 67–90.
[15] Weiss, L. and Wolfowitz, J. (1973). Maximum likelihood estimation of a translation of a truncated distribution. *Ann. Statist.* **1** 944–947.

IMS Lecture Notes–Monograph Series
Time Series and Related Topics
Vol. 52 (2006) 284–294

DOI: 10.1214/074921706000001111

Some results on the Gittins index for a normal reward process

Yi-Ching Yao[1]

Academia Sinica

Abstract: We consider the Gittins index for a normal distribution with unknown mean θ and known variance where θ has a normal prior. In addition to presenting some monotonicity properties of the Gittins index, we derive an approximation to the Gittins index by embedding the (discrete-time) normal setting into the continuous-time Wiener process setting in which the Gittins index is determined by the stopping boundary for an optimal stopping problem. By an application of Chernoff's continuity correction in optimal stopping, the approximation includes a correction term which accounts for the difference between the discrete and continuous-time stopping boundaries. Numerical results are also given to assess the performance of this simple approximation.

1. Introduction

The classical multi-armed bandit problem is concerned with sequential design of adaptive sampling from k statistical populations with distribution functions F_{θ_i}, $i = 1, \ldots, k$ $(k \geq 2)$ where θ_i denotes the unknown parameter of the ith population. Specifically, the objective is to sample $Y_1, Y_2, \ldots$ sequentially from the k populations so as to maximize the expected total discounted reward

$$E_\pi E_{\theta_1,\ldots,\theta_k}\left(\sum_{j=1}^{\infty}\gamma_j Y_j\right) = \int E_{\theta_1,\ldots,\theta_k}\left(\sum_{j=1}^{\infty}\gamma_j Y_j\right) d\pi(\theta_1,\ldots,\theta_k),$$

where π is the prior distribution of $(\theta_1, \ldots, \theta_k)$ and $\{\gamma_j\}$ is a (deterministic) discount sequence. The two most important types of discount sequence are uniform discounting with finite horizon $N > 0$ (i.e. $\gamma_j = 1$ for $j \leq N$ and $\gamma_j = 0$ for $j > N$) and geometric discounting with discount factor $0 < \beta < 1$ (i.e. $\gamma_j = \beta^{j-1}, j = 1, 2, \ldots$). While in general the optimal allocation rule can only be characterized via the dynamic programming equations which admit no general closed-form solutions, Gittins and Jones [13] showed that under geometric discounting, when the prior distribution is a product measure $d\pi(\theta_1, \ldots, \theta_k) = d\pi_1(\theta_1) \times \cdots \times d\pi_k(\theta_k)$, the optimal allocation rule is to sample at each stage from the population with the greatest (current) Gittins index. See also [11] and [21].

For a population with distribution function F_θ and (current) prior distribution $\pi(\theta)$ of the unknown parameter θ, the Gittins index is defined by

$$\lambda(\pi,\beta) = \sup_{\xi\geq 1}\left[E_\pi E_\theta\left(\sum_{n=1}^{\xi}\beta^{n-1}X_n\right)\Big/ E_\pi E_\theta\left(\sum_{n=1}^{\xi}\beta^{n-1}\right)\right] \tag{1}$$

[1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, e-mail: yao@stat.sinica.edu.tw

AMS 2000 subject classifications: primary 60G40; secondary 90C39.

Keywords and phrases: Chernoff's continuity correction, dynamic allocation index, multi-armed bandit, optimal stopping, Brownian motion.

where the supremum is taken over all (integer-valued) stopping times $\xi \geq 1$ and $X_1, X_2, \ldots$ are (conditionally) iid with common distribution function F_θ (given θ). Equivalently, $\lambda(\pi, \beta)$ is the infimum of the set of solutions λ of the equation

$$\frac{\lambda}{1-\beta} = \sup_{\xi \geq 0} E_\pi E_\theta \left[\sum_{n=1}^{\xi} \beta^{n-1} X_n + \beta^\xi \frac{\lambda}{1-\beta} \right]. \tag{2}$$

In [12], computational methods for calculating Gittins indices are described and applied to the normal, Bernoulli and negative exponential families with conjugate priors, which involve using backward induction to approximate the right-hand side of (2) with the supremum over $\xi \geq 0$ replaced by the supremum over $0 \leq \xi \leq N$ for some large horizon N. For β close to 1, such computational methods become time consuming as a very large horizon N is required to yield an accurate approximation. Thus it will be useful to have accurate analytic approximations to Gittins indices especially for β close to 1.

In this paper, we consider the normal case with unknown mean θ and known variance where θ has a normal (conjugate) prior. Section 2 presents some monotonicity properties of the Gittins index. In particular, it is shown that the Gittins index is a nondecreasing function of the prior variance. In Section 3, a corrected diffusion approximation to the Gittins index is derived by embedding the (discrete-time) normal setting into the continuous-time Wiener process setting in which the Gittins index is determined by the stopping boundary for an optimal stopping problem (first introduced in [2]). By an application of Chernoff's continuity correction, the approximation includes a correction term which accounts for the difference between the discrete and continuous-time stopping boundaries. Numerical results are also given to assess the performance of this simple approximation. To prepare for the derivations, Sections 3.1 and 3.2 briefly review, respectively, some properties of the Gittins index for a Wiener process and Chernoff's continuity correction in optimal stopping.

The monograph of Gittins [12] provides a comprehensive theory of dynamic allocation indices and explores the class of problems whose optimal solutions can be characterized by dynamic allocation indices. On the other hand, Lai [17] and Chang and Lai [6] have proposed simple index-type adaptive allocation rules that are asymptotically optimal in both the Bayes and frequentist senses either as $N \to \infty$ (under uniform discounting) or as $\beta \to 1$ (under geometric discounting). Brezzi and Lai [5] have recently refined and modified these adaptive allocation rules in the presence of switching costs, while Hu and Wei [15] have constructed asymptotically optimal adaptive allocation rules subject to the irreversibility constraint. Various applications of the theory of multi-armed bandits can be found in sequential clinical trials, market pricing, labor markets and search problems; see e.g. [1, 8, 16, 19, 20].

2. Some monotonicity properties of the Gittins index for a normal reward process

In this section, we consider the case that $X_1, X_2, \ldots$ are (conditionally) iid $N(\theta, \sigma^2)$, the unknown mean θ has a prior $\pi = N(u, v)$ and the variance σ^2 is known. The Gittins index is denoted by $\lambda(u, v, \sigma^2, \beta)$. By location and scale equivariance properties (cf. [12], Section 6.4),

$$\lambda(u, v, \sigma^2, \beta) = u + r\,\lambda(0,\, v/r^2,\, \sigma^2/r^2,\, \beta) \tag{3}$$

for $r > 0$.

Lemma 1. *The Gittins index $\lambda\,(u, v, \sigma^2, \beta)$ is nonincreasing in σ^2.*

Proof. We prove the lemma by a simple randomization argument. Fix $0 < \sigma_1^2 < \sigma_2^2$. Let $X_1, X_2, \dots$ be (conditionally) iid $N(\theta, \sigma_1^2)$ given θ, which is assumed to have a prior $\pi = N(u, v)$. Let $\epsilon_1, \epsilon_2, \dots$ be iid $N(0,\, \sigma_2^2 - \sigma_1^2)$ (independent of the X_i). Then $X_1' = X_1 + \epsilon_1$, $X_2' = X_2 + \epsilon_2, \dots$ are (conditionally) iid $N(\theta, \sigma_2^2)$ given θ. For any stopping time $\xi' \geq 1$ with respect to the filtration $\mathcal{F}'$ generated by $X_1', X_2', \dots$, we have

$$\begin{aligned} E_\pi\, E_\theta \sum_{n=1}^{\xi'} \beta^{n-1} X_n' &= E_\pi\, E_\theta \sum_{n=1}^{\xi'} \beta^{n-1} X_n + E_\pi\, E_\theta \sum_{n=1}^{\infty} \beta^{n-1}\, \epsilon_n\, 1_{\{\xi' \geq n\}} \\ &= E_\pi\, E_\theta \sum_{n=1}^{\xi'} \beta^{n-1} X_n\,. \end{aligned}$$

Since every stopping time ξ' with respect to $\mathcal{F}'$ may be viewed as a randomized stopping time with respect to $\mathcal{F}$ (the filtration generated by $X_1, X_2, \dots$), it follows that

$$\begin{aligned} \lambda(u,\, v,\, \sigma_2^2, \beta) \;&=\; \sup_{\xi' \geq 1} E_\pi\, E_\theta\Big(\sum_{n=1}^{\xi'} \beta^{n-1} X_n'\Big) \Big/ E_\pi\, E_\theta\Big(\sum_{n=1}^{\xi'} \beta^{n-1}\Big) \\ &\leq\; \sup_{\xi \geq 1} E_\pi\, E_\theta\Big(\sum_{n=1}^{\xi} \beta^{n-1} X_n\Big) \Big/ E_\pi\, E_\theta\Big(\sum_{n=1}^{\xi} \beta^{n-1}\Big) \\ &=\; \lambda(u,\, v,\, \sigma_1^2, \beta)\,, \end{aligned}$$

completing the proof. □

Theorem 1. *$\lambda\,(0,\, v,\, \sigma^2,\, \beta)/\sqrt{v}$ is nondecreasing in v.*

Proof. For fixed $0 < v_2 < v_1$, it follows from (3) and Lemma 1 that

$$\begin{aligned} \lambda(0,\, v_2,\, \sigma^2, \beta) &= \sqrt{v_2/v_1}\; \lambda(0,\, v_1,\, \sigma^2\, v_1/v_2,\, \beta) \\ &\leq \sqrt{v_2/v_1}\; \lambda(0,\, v_1,\, \sigma^2,\, \beta)\,, \end{aligned}$$

completing the proof. □

Corollary 1. *$\lambda(u,\, v,\, \sigma^2,\, \beta) = u + \lambda(0,\, v,\, \sigma^2,\, \beta)$ is nondecreasing in u and v.*

Remark 1. For the Wiener process setting, Bather [2] proved a result analogous to Theorem 1 (see (7) and (8) below).

Remark 2. For a normal two-armed bandit in which the means of arms 1 and 2 have independent normal priors $N(u_1, v_1)$ and $N(u_2, v_2)$ and their variances are known and equal, it follows from Corollary 1 that under geometric discounting, it is optimal to pull arm 1 initially if $u_1 \geq u_2$ and $v_1 \geq v_2$. It seems natural to conjecture that the same also holds under uniform discounting. Note that Berry [3] made a similar conjecture regarding a Bernoulli two-armed bandit, which has not been resolved (cf. [4], Section 7.3).

Remark 3. Along the lines of the proof of Theorem 1, it can be readily shown that

$$\lambda(0,\, v_1,\, \sigma_1^2,\, \beta)/\sqrt{v_1} \geq \lambda(0,\, v_2,\, \sigma_2^2,\, \beta)/\sqrt{v_2}$$

if $v_1 \geq v_2$ and $v_1/\sigma_1^2 \geq v_2/\sigma_2^2$. Note that for a normal distribution $N(\theta, \sigma^2)$ where θ has a normal prior $N(0, v)$, v/σ^2 may be referred to as the signal-to-noise ratio since v is the second moment of the "signal" θ.

3. Corrected diffusion approximation to the Gittins index for a normal reward process

In Section 3.3, we derive an approximation to the Gittins index for a normal distribution whose mean is assumed to have a normal prior. To prepare for the derivations, we briefly review, in Sections 3.1 and 3.2, some properties of the Gittins index for a Wiener process and Chernoff's continuity correction in optimal stopping.

3.1. Properties of the Gittins index for a Wiener process

Bather [2] showed that for a Wiener process $\{W(t), t \geq 0\}$ with drift coefficient θ which has a normal prior $N(u_0, v_0)$, the Gittins index $\lambda^*(u_0, v_0, c)$ can be determined by the solution to an optimal stopping problem (to be described below) where $c > 0$ denotes the discount rate in continuous time (see also [6] and Section 6.6 of [12]). Here $\lambda^*(u_0, v_0, c)$ is defined as the infimum of the set of solutions λ of the equation (cf. (2))

$$
\begin{aligned}
\lambda \int_0^\infty e^{-c\,t} dt \;&=\; \sup_{\tau \geq 0} E_\pi \, E_\theta \left[\int_0^\tau e^{-ct} dW(t) + \lambda \int_\tau^\infty e^{-ct} dt \right] \\
&=\; \sup_{\tau \geq 0} E_\pi \left[\int_0^\tau \theta e^{-ct} dt + \lambda \int_\tau^\infty e^{-ct} dt \right] \\
&=\; \sup_{\tau \geq 0} E_\pi \left[\int_0^\tau u(t)\, e^{-ct} dt + \lambda \int_\tau^\infty e^{-ct} dt \right] \\
&=\; \sup_{\tau \geq 0} E_\pi \Big[c^{-1} u_0 - c^{-1} \Big(u(\tau) - \lambda \Big) e^{-c\tau} \Big] \,,
\end{aligned}
\tag{4}
$$

where the supremum is taken over all (real-valued) stopping times $\tau \geq 0$, $\pi = N(u_0, v_0)$ is the prior distribution of θ, and $u(t)$ is the posterior mean of θ, i.e.

$$
u\,(t) = E_\pi \Big[\, \theta \mid W(s),\, 0 \leq s \leq t \,\Big] = \frac{v_0^{-1} u_0 + W(t)}{v_0^{-1} + t} \,. \tag{5}
$$

The last equality in (4) follows from integration by parts along with the fact that a simple change of time transforms u into standard Brownian motion, cf. $Y(v)$ below.

Define

$$
\begin{aligned}
&v = v(t) = (v_0^{-1} + t)^{-1} \text{ (the posterior variance)}, \;\; s = v/c \,, \\
&Y(v) = u_0 - u(t), \;\; \text{and} \;\; Z(s) = Y(cs)/\sqrt{c} \,.
\end{aligned}
$$

It can be readily shown that $\{Y(v),\, 0 < v \leq v_0\}$ is standard Brownian motion ($Y(v_0) = 0$) in the $-v$ scale and $\{Z(s),\, 0 < s \leq s_0\}$ ($s_0 = v_0/c$) is standard Brownian motion ($Z(s_0) = 0$) in the $-s$ scale. Letting $z_0 = (\lambda - u_0)/\sqrt{c}$, it follows that (4) is equivalent to

$$
z_0\, e^{-1/s_0} = \sup_{0 < S \leq s_0} E \left[\, \{Z(S) + z_0\} \; e^{-1/S} \, \right] \tag{6}
$$

in the sense that λ is a solution of (4) if and only if $z_0 = (\lambda - u_0)/\sqrt{c}$ is a solution of (6), where the supremum on the right-hand side of (6) is taken over all stopping

times $0 < S \le s_0$ (in the $-s$ scale). It is more convenient to remove the restriction of $Z(s_0) = 0$ and rewrite (6) as

$$z_0\, e^{-1/s_0} = \sup_{0<S\le s_0} E\Big[\, Z(S)\, e^{-1/S} \mid Z(s_0) = z_0 \,\Big]\,. \tag{6'}$$

For the optimal stopping problem with payoff function $g(z,s) = ze^{-1/s}$ on the right-hand side of (6′), it is easily shown that the continuation region is of the form $\{(z,s) : z < b(s)\}$ where $b(s) > 0$ is the optimal stopping boundary. Since z_0 is a solution of (6′) if and only if (z_0, s_0) is in the stopping region (i.e. $z_0 \ge b(s_0)$), it follows that $\lambda^*(u_0, v_0, c)$, the infimum of the set of solutions λ of the equation (4), satisfies $b(s_0) = \big(\lambda^*(u_0,v_0,c) - u_0\big)/\sqrt{c}$, i.e.

$$\lambda^*(u_0,v_0,c) = u_0 + \sqrt{c}\, b(s_0) = u_0 + \sqrt{c}\, b(v_0/c)\,. \tag{7}$$

Bather [2] showed that

$$b(s)/\sqrt{s} \text{ is a nondecreasing function of } s\,, \tag{8}$$

$$b(s) \le s/\sqrt{2} \text{ for all } s > 0\,, \text{ and } \lim_{s\to 0} b(s)/s = 1/\sqrt{2}\,, \tag{9}$$

while Chang and Lai [6] derived the asymptotic expansion as $s \to \infty$

$$b(s) = \left\{2s\Big[\log s - \frac{1}{2}\log\log s - \frac{1}{2}\log 16\pi + o(1)\Big]\right\}^{1/2}. \tag{10}$$

Based on (8)–(10) together with extensive numerical work (involving corrected Bernoulli random walk approximations for Brownian motion), Brezzi and Lai [5] have suggested the following closed-form approximation $\Psi(s)$ to $b(s)/\sqrt{s}$

$$\frac{b(s)}{\sqrt{s}} \approx \Psi(s) = \begin{cases} \sqrt{s/2} & \text{for } s \le 0.2\,, \\ 0.49 - 0.11\, s^{-1/2} & \text{for } 0.2 < s \le 1\,, \\ 0.63 - 0.26\, s^{-1/2} & \text{for } 1 < s \le 5\,, \\ 0.77 - 0.58\, s^{-1/2} & \text{for } 5 < s \le 15\,, \\ \Big\{2\log s - \log\log s - \log 16\pi\Big\}^{1/2} & \text{for } s > 15\,. \end{cases} \tag{11}$$

3.2. Chernoff's continuity correction in optimal stopping

In his pioneering work, Chernoff [7] studied the relationship between the solutions of the discrete and continuous-time versions of the problem of testing sequentially the sign of the mean of a normal distribution. His result may be stated more generally as follows. Let $\{B(t)\}$ be standard Brownian motion and let $g(x,t)$ be a smooth payoff function for $t \le T$ (horizon) for which the continuation region is of the form $\{(x,t) : x < b(t)\}$. Consider a constrained optimal stopping problem where stopping is permitted only at $n\delta$, $n = 1, 2, \ldots$ where δ is a given (small) positive number. Suppose that there exist stopping boundary points $b_\delta(n\delta)$, $n = 1, 2, \ldots$ such that starting from $B(n_0\delta) = x_0$ for any given n_0 and x_0, the optimal stopping rule is to stop at the first $n \ge n_0$ at which $B(n\delta) \ge b_\delta(n\delta)$. So $b_\delta(n\delta)$ (or $b(t)$, resp.) is the

discrete-time (or continuous-time, resp.) stopping boundary for the constrained (or unconstrained, resp.) optimal stopping problem. Then for fixed $t < T$, we have

$$b_\delta(t) = b(t) - \rho\sqrt{\delta} + o(\sqrt{\delta}) \quad \text{as } \delta \to 0\,, \tag{12}$$

where $b_\delta(t) = b_\delta(\,[\,t/\delta\,]\,\delta\,)$, $\rho = ES^2_{\tau_+}\,/2ES_{\tau_+} \approx 0.583$, $\tau_+ = \inf\{\,n : S_n > 0\}$, $S_n = X_1 + \cdots + X_n$, and the X_i are iid $N(0,1)$.

Chernoff [7] derived (12) by relating the original problem to an associated stopping problem in which there is a horizon at $t = 0$ and the payoff function is $g(x,t) = -t + x^2\,1_{\{x<0,\,t=0\}}$, $t \le 0$. For the associated stopping problem, stopping is permitted at $0, -1, -2, \ldots$, and there exist stopping boundary points $b_{-1} > b_{-2} > \cdots$ such that starting from (x_0, n_0) with $n_0 < 0$, the optimal stopping rule is to stop at the first $n_0 \le n \le 0$ at which

$$x_0 + X_1 + \cdots + X_{n-n_0} \ge b_n \quad (b_0 = -\infty).$$

Chernoff [7] and subsequently Chernoff and Petkau [9] and Hogan [14] showed that

$$\lim_{n\to-\infty} b_n = -ES^2_{\tau_+}/2\,ES_{\tau_+}$$

for normal, Bernoulli and general X (with finite fourth moment), respectively. Recently, under mild growth conditions on g, Lai, Yao and AitSahlia [18] have proved (12) when the Brownian motion process is replaced by a general random walk in the constrained problem.

3.3. Approximating the Gittins index for a normal reward process

In this subsection, we consider the case that $X_1, X_2, \ldots$ are (conditionally) iid $N(\theta, \sigma^2)$ and the unknown mean θ has a prior $\pi = N(u_0, v_0)$. Without loss of generality, we assume $\sigma^2 = 1$. For notational simplicity, the Gittins index $\lambda(u_0, v_0, 1, \beta)$ will be abbreviated to $\lambda(u_0, v_0, \beta)$. Recall that $\lambda(u_0, v_0, \beta)$ is the infimum of the set of solutions λ of the equation (2). As in Section 3.1, let $\{W(t), t \ge 0\}$ be a Wiener process with drift coefficient θ which has a normal prior $N(u_0, v_0)$. Noting that $(X_1, X_2, \ldots)$ and $(W(1), W(2) - W(1), \ldots)$ have the same joint distribution, we can rewrite (2) as

$$\begin{aligned}
\frac{\lambda}{1-\beta} &= \sup_{\xi\,\ge\,0} E_\pi\, E_\theta\Big[\sum_{n=1}^{\xi} \beta^{n-1}\Big(W\big(n\big) - W\big(n-1\big)\Big) + \beta^\xi \frac{\lambda}{1-\beta}\Big] \\
&= \sup_{\xi\,\ge\,0} E_\pi\Big[\sum_{n=1}^{\xi} \beta^{n-1}\, u\big(n-1\big) + \beta^\xi \frac{\lambda}{1-\beta}\Big] \\
&= \sup_{\xi\,\ge\,0} E_\pi\Big[\frac{c}{1-\beta}\int_0^{\xi} u\big(t\big)\, e^{-ct} dt + \beta^\xi \frac{\lambda}{1-\beta}\Big] \\
&= \frac{1}{1-\beta} \sup_{\xi\,\ge\,0} E_\pi\Big[u_0 - \big(\,u(\xi) - \lambda\,\big)\, e^{-c\xi}\Big],
\end{aligned}$$

where $u(t)$ is given in (5), $c = -\log\beta$ and the third equality follows since

$$\begin{aligned}
&E\Big[\frac{c}{1-\beta}\int_{n-1}^{n} u(t)\, e^{-ct} 1_{\{\xi\ge n\}}\, dt \mid W(s),\, 0 \le s \le n-1\Big] \\
&\quad = \frac{c}{1-\beta}\, 1_{\{\xi\ge n\}} \int_{n-1}^{n} u(n-1)\, e^{-ct}\, dt \;=\; \beta^{n-1}\, u(n-1) 1_{\{\xi\ge n\}}\,.
\end{aligned}$$

With the notation introduced in Section 3.1, we can further rewrite (2) as

$$\lambda - u_0 = \sup_{V \in \{v_0/(1+v_0 n)\,,\, n=0,1,\ldots\}} E\Big[\Big(\lambda - u_0 + Y(V) \Big)\, e^{-cV^{-1} + c\, v_0^{-1}} \Big]$$

where the supremum is taken over all stopping times V taking values in $\{ v_0/(1 + v_0 n),\, n = 0, 1, \ldots \}$. In terms of Brownian motion $Z(s)$ in the $-s$ scale, (2) is equivalent to

$$z_0 e^{-1/s_0} = \sup_{S \,\in\, \{c^{-1} v_0/(1+v_0 n)\,,\, n=0,1,\ldots\}} E\Big[Z(S)\, e^{-1/S} \mid Z(s_0) = z_0 \Big] \tag{13}$$

where $z_0 = (\lambda - u_0)/\sqrt{c}$, $s_0 = v_0/c$ and the supremum is taken over all stopping times S taking values in $\{c^{-1} v_0/(1 + v_0 n)\,,\, n = 0, 1, \ldots\}$.

For the constrained optimal stopping problem on the right-hand side of (13), there exist optimal stopping boundary points $b_{v_0}\big(c^{-1} v_0/(1+v_0 n)\big)$, $n = 0, 1, \ldots$ such that the optimal stopping rule is to stop at the first n at which $Z\big(c^{-1} v_0/(1+v_0 n)\big) \geq b_{v_0}\big(c^{-1} v_0/(1 + v_0 n)\big)$. So $b_{v_0}(v_0/c)$ is the infimum of the set of solutions z_0 of the equation (13). It then follows that

$$\lambda\big(u_0, v_0, \beta\big) = u_0 + \sqrt{c}\, b_{v_0}(v_0/c)\,. \tag{14}$$

Since in the constrained optimal stopping problem the permissible stopping time points $c^{-1} v_0/(1 + v_0 n)$, $n = 0, 1, 2, \ldots$ are not equally spaced, there is no rigorous justification for applying (12) to relate the discrete and continuous-time stopping boundaries $b_{v_0}(t)$ and $b(t)$ for the constrained and unconstrained problems. However, it can be argued heuristically that (12) applies when the spacing between many successive permissible stopping time points is approximately constant (cf. bottom of page 47 in [10]). Thus we arrive at the following approximation

$$b_{v_0}(v_0/c) \approx b(v_0/c) - 0.583\sqrt{\delta} \tag{15}$$

where

$$\delta = \frac{c^{-1} v_0}{1 + v_0 \cdot 0} - \frac{c^{-1} v_0}{1 + v_0 \cdot 1} = \frac{c^{-1} v_0^2}{1 + v_0}\,, \tag{16}$$

provided that v_0/c is bounded away from 0 (the horizon of the optimal stopping problem) and $\delta \approx \frac{c^{-1} v_0}{1+v_0 n} - \frac{c^{-1} v_0}{1+v_0(n+1)}$ for many (small to moderate) n's. That is, we expect the approximation (15) to be reasonably good if v_0 is small and v_0/c is not too close to 0. It follows from (14), (15), (16) and (11) that

$$\begin{aligned} \lambda\big(u_0, v_0, \beta\big) &\approx u_0 + \sqrt{c}\, b(v_0/c) - 0.583\, v_0/\sqrt{1 + v_0} \\ &\approx u_0 + \sqrt{v_0}\, \Psi(v_0/c) - 0.583\, v_0/\sqrt{1 + v_0}\,. \end{aligned} \tag{17}$$

Note that the continuation region for the constrained problem must be contained in the continuation region for the unconstrained problem, so that $b_{v_0}(v_0/c) < b(v_0/c)$. Thus the uncorrected diffusion approximation $u_0 + \sqrt{c} b(v_0/c)$ overestimates $\lambda(u_0, v_0, \beta) = u_0 + \sqrt{c}\, b_{v_0}(v_0/c)$, which is recorded in the following theorem.

Theorem 2. $\lambda(u_0, v_0, \beta) < u_0 + \sqrt{c}\, b(v_0/c)$ *where* $c = \log \beta^{-1}$.

A related upper bound for $\lambda(u_0, v_0, \beta)$ is given in Theorem 6.28 of Gittins [12], which states, in our notation, that

$$\lambda(u_0, v_0, \beta) < u_0 + \sqrt{1-\beta}\ \ b\big(v_0/(1-\beta)\big)\,. \tag{18}$$

Since $b(s)/\sqrt{s}$ is nondecreasing in s by (8) and since $c = \log \beta^{-1} > 1-\beta$, we have

$$\sqrt{c}\, b(v_0/c) \le \sqrt{1-\beta}\ \ b\big(v_0/(1-\beta)\big)\,,$$

so that the upper bound given in Theorem 2 is sharper than (18).

In the approximation (15), the correction term $0.583\sqrt{\delta}$ with δ given in (16) appears to be a little too large since the spacing between successive permissible stopping time points $\frac{c^{-1}v_0}{1+v_0 n} - \frac{c^{-1}v_0}{1+v_0(n+1)}$ is strictly less than δ for $n \ge 1$. To compensate for this overcorrection, we propose (in view of (9)) to replace $b(v_0/c)$ by $(v_0/c)/\sqrt{2}$ in (15), resulting in the following simple approximation

$$\lambda\big(u_0, v_0, \beta\big) \approx u_0 + v_0/\sqrt{2c} - 0.583\, v_0/\sqrt{1+v_0}\,. \tag{19}$$

Note that (19) agrees with (17) for $v_0/c \le 0.2$ in view of (11).

In his Table 1, Gittins [12] tabulates $n(1-\beta)^{1/2}\,\lambda(\,0, n^{-1}, \beta\,)$ for various values of n and β. Our Table 1 compares $n(1-\beta)^{1/2}\,\lambda(\,0, n^{-1}, \beta\,)$ with the corrected and uncorrected approximations (based on (17) and (19))

$$\begin{aligned}
&\text{(CA)} \quad n(1-\beta)^{\frac12}\left[\frac{1}{\sqrt{n}}\,\Psi\big(\tfrac{1}{nc}\big) - \frac{0.583\,n^{-1}}{\sqrt{1+n^{-1}}}\right] \\
&\qquad\qquad = (1-\beta)^{\frac12}\left[\sqrt{n}\,\Psi\big(\tfrac{1}{nc}\big) - \frac{0.583}{\sqrt{1+n^{-1}}}\right], \\
&\text{(UA)} \quad n(1-\beta)^{\frac12}\,\frac{1}{\sqrt{n}}\,\Psi\big(\tfrac{1}{nc}\big) = (1-\beta)^{\frac12}\,\sqrt{n}\,\Psi\big(\tfrac{1}{nc}\big)\,, \\
&\text{(CA}')\quad (1-\beta)^{\frac12}\left[\frac{1}{\sqrt{2c}} - \frac{0.583}{\sqrt{1+n^{-1}}}\right], \\
&\text{(UA}')\quad \sqrt{(1-\beta)/(2c)}\,.
\end{aligned}$$

Remark 4. As explained earlier, the uncorrected approximations have positive bias due to overestimation. The corrected approximations are reasonably accurate for moderate to large n and for large β. For moderate n, (CA) (or (CA′), resp.) tends to underestimate (or overestimate, resp.) $n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$. This observation naturally leads to approximating $n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$ by the average of (CA) and (CA′), which is also included in Table 1. Overall, $[\,\text{(CA)} + \text{(CA}')\,]/2$ has the best performance, while (CA′) is better than (CA) except for small n and large β.

Remark 5. Table 1 of Gittins [12] suggests that $n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$ is increasing in n. For $\beta = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95$, Gittins has numerically estimated $\lim_{n\to\infty} n(1-\beta)^{1/2}\lambda(0, n^{-1}, \beta)$. These numbers are compared in Table 2 with the limits $(1-\beta)^{1/2}[\,(2c)^{-1/2} - 0.583\,]$ (or $(1-\beta)^{1/2}/(2c)^{1/2}$, resp.) obtained from the corrected approximations (CA) and (CA′) (or uncorrected approximations (UA) and (UA′), resp.) as $n \to \infty$. It should be noted that the heuristic justification for the corrected approximations requires $v_0/c = 1/(n\,c)$ not to be very close to 0.

TABLE 1
Gittins indices and approximations
($\beta = 0.5, 0.7, 0.9, 0.95, 0.99, 0.995$)

	n				
	10	50	100	500	1000
$\beta = 0.5$					
$n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$	0.211	0.224	0.226	0.227	0.227
$[\,(\mathrm{CA}) + (\mathrm{CA}')\,]/2$	0.208	0.192	0.190	0.189	0.189
(CA)	0.208	0.192	0.190	0.189	0.189
(CA′)	0.208	0.192	0.190	0.189	0.189
(UA)	0.601	0.601	0.601	0.601	0.601
(UA′)	0.601	0.601	0.601	0.601	0.601
$\beta = 0.7$					
$n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$	0.311	0.337	0.341	0.344	0.345
$[\,(\mathrm{CA}) + (\mathrm{CA}')\,]/2$	0.264	0.332	0.331	0.329	0.329
(CA)	0.184	0.332	0.331	0.329	0.329
(CA′)	0.344	0.332	0.331	0.329	0.329
(UA)	0.489	0.648	0.648	0.648	0.648
(UA′)	0.648	0.648	0.648	0.648	0.648
$\beta = 0.9$					
$n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$	0.415	0.480	0.493	0.504	0.506
$[\,(\mathrm{CA}) + (\mathrm{CA}')\,]/2$	0.357	0.506	0.505	0.505	0.505
(CA)	0.201	0.506	0.505	0.505	0.505
(CA′)	0.513	0.506	0.505	0.505	0.505
(UA)	0.377	0.689	0.689	0.689	0.689
(UA′)	0.689	0.689	0.689	0.689	0.689
$\beta = 0.95$					
$n(1-\beta)^{1/2}\lambda(0, n^{-1}, \beta)$	0.425	0.519	0.540	0.562	0.566
$[\,(\mathrm{CA}) + (\mathrm{CA}')\,]/2$	0.382	0.468	0.568	0.568	0.568
(CA)	0.190	0.367	0.568	0.568	0.568
(CA′)	0.574	0.569	0.568	0.568	0.568
(UA)	0.314	0.496	0.698	0.698	0.698
(UA′)	0.698	0.698	0.698	0.698	0.698
$\beta = 0.99$					
$n(1-\beta)^{1/2}\lambda(0, n^{-1}, \beta)$	0.353	0.499	0.549	0.618	0.633
$[\,(\mathrm{CA}) + (\mathrm{CA}')\,]/2$	0.390	0.453	0.485	0.647	0.647
(CA)	0.130	0.257	0.322	0.647	0.647
(CA′)	0.650	0.648	0.647	0.647	0.647
(UA)	0.185	0.315	0.380	0.705	0.705
(UA′)	0.705	0.705	0.705	0.705	0.705
$\beta = 0.995$					
$n(1-\beta)^{1/2}\lambda(0, n^{-1}, \beta)$	0.304	0.457	0.516	0.614	0.638
$[\,(\mathrm{CA}) + (\mathrm{CA}')\,]/2$	0.424	0.437	0.470	0.562	0.665
(CA)	0.181	0.209	0.274	0.458	0.665
(CA′)	0.667	0.665	0.665	0.665	0.665
(UA)	0.221	0.250	0.315	0.499	0.706
(UA′)	0.706	0.706	0.706	0.706	0.706

Remark 6. Brezzi and Lai [5] have proposed a simple approximation to Gittins indices for general distributions which is justified by making use of the functional central limit theorem as $\beta \to 1$. For Bernoulli distributions (with beta conjugate priors), their approximation provides fairly accurate results. When applied to normal distributions, their approximation reduces to the uncorrected approximation (UA). It will be of great interest to see whether and how Chernoff's continuity correction can apply to approximate Gittins indices for nonnormal distributions.

TABLE 2
The limits of Gittins indices and approximations

β	$\lim_{n\to\infty} n(1-\beta)^{1/2}\,\lambda(0, n^{-1}, \beta)$	(CA) and (CA′)	(UA) and (UA′)
0.5	0.227	0.189	0.601
0.6	0.283	0.257	0.626
0.7	0.345	0.329	0.648
0.8	0.417	0.409	0.669
0.9	0.509	0.505	0.689
0.95	0.583	0.568	0.698

References

[1] BANKS, J. S. AND SUNDARAM, R. K. (1992). Denumerable-armed bandits. *Econometrica* **60** 1071–1096.

[2] BATHER, J. A. (1983). Optimal stopping of Brownian motion: A comparison technique. In *Recent Advances in Statistics*, ed. J. Rustagi et al. Academic Press, New York, 19–50.

[3] BERRY, D. A. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897.

[4] BERRY, D. A. AND FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London.

[5] BREZZI, M. AND LAI, T. L. (2002). Optimal learning and experimentation in bandit problems. *J. Economic Dynamics & Control* **27** 87–108.

[6] CHANG, F. AND LAI, T. L. (1987). Optimal stopping and dynamic allocation. *Adv. Appl. Probab.* **19** 829–853.

[7] CHERNOFF, H. (1965). Sequential tests for the mean of a normal distribution IV (discrete case). *Ann. Math. Statist.* **36** 55–68.

[8] CHERNOFF, H. (1967). Sequential models for clinical trials. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 805–812. University of California Press.

[9] CHERNOFF, H. AND PETKAU, A. J. (1976). An optimal stopping problem for sums of dichotomous random variables. *Ann. Probab.* **4** 875–889.

[10] CHERNOFF, H. AND PETKAU, A. J. (1986). Numerical solutions for Bayes sequential decision problems. *SIAM J. Scient. Statist. Comput.* **7** 46–59.

[11] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Royal Stat. Soc.* Series B **41** 148–177.

[12] GITTINS, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. Wiley, New York.

[13] GITTINS, J. C. AND JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, ed. J. Gani et al. North-Holland, Amsterdam, pp. 241–266.

[14] HOGAN, M. (1986). Comments on a problem of Chernoff and Petkau. *Ann. Probab.* **14** 1058–1063.

[15] HU, I. AND WEI, C. Z. (1989). Irreversible adaptive allocation rules. *Ann. Statist.* **17** 801–823.

[16] JOVANOVIC, B. (1979). Job matching and the theory of turnover. *J. Political Econ.* **87** 972–990.

[17] LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15** 1091–1114.

[18] LAI, T. L., YAO, Y.-C. AND AITSAHLIA, F. (2005). Corrected random walk approximations to free boundary problems in optimal stopping. Technical report, Department of Statistics, Stanford University.

[19] Mortensen, D. (1985). Job-search and labor market analysis. In *Handbook of Labor Economics* Vol. 2, O. Ashenfelter and R. Layard, ed. North-Holland, Amsterdam, 849–919.
[20] Rothschild, M. (1974). A two-armed bandit theory of market pricing. *J. Econ. Theory* **9** 185-202.
[21] Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. Royal Stat. Soc.* Series B **42** 143–149.