Computer-Aided Econometrics

edited by
David E. A. Giles
University of Victoria
Victoria, British Columbia, Canada
STATISTICS: Textbooks and Monographs  
D. B. Owen  
_Founding Editor, 1972–1991_  

**Associate Editors**  
- **Statistical Computing/Nonparametric Statistics**  
  Professor William R. Schucany  
  *University of Michigan*  
- **Multivariate Analysis**  
  Professor Anant M. Kshirsagar  
  *University of Michigan*  
- **Probability**  
  Professor Marcel F. Neuts  
  *University of Arizona*  
- **Quality Control/Reliability**  
  Professor Edward G. Schilling  
  *Rochester Institute of Technology*  

**Editorial Board**  
- **Applied Probability**  
  Dr. Paul R. Garvey  
  *The MITRE Corporation*  
- **Statistical Distributions**  
  Professor N. Balakrishnan  
  *McMaster University*  
- **Economic Statistics**  
  Professor David E. A. Giles  
  *University of Victoria*  
- **Statistical Process Improvement**  
  Professor G. Geoffrey Vining  
  *Virginia Polytechnic Institute*  
- **Experimental Designs**  
  Mr. Thomas B. Barker  
  *Rochester Institute of Technology*  
- **Stochastic Processes**  
  Professor V. Lakshmikantham  
  *Florida Institute of Technology*  
- **Multivariate Analysis**  
  Professor Subir Ghosh  
  *University of California–Riverside*  
- **Survey Sampling**  
  Professor Lynne Stokes  
  *Southern Methodist University*  
- **Time Series**  
  Sastry G. Pantula  
  *North Carolina State University*
1. The Generalized Jackknife Statistic, H. L. Gray and W. R. Schucany
2. Multivariate Analysis, Anant M. Kshirsagar
5. Design of Experiments: A Realistic Approach, Virgil L. Anderson and Robert A. McLean
6. Statistical and Mathematical Aspects of Pollution Problems, John W. Pratt
7. Introduction to Probability and Statistics (in two parts), Part I: Probability; Part II: Statistics, Narayan C. Giri
8. Statistical Theory of the Analysis of Experimental Designs, J. Ogawa
9. Statistical Techniques in Simulation (in two parts), Jack P. C. Kleijnen
10. Data Quality Control and Editing, Joseph I. Naus
11. Cost of Living Index Numbers: Practice, Precision, and Theory, Kali S. Banerjee
13. The Search for Oil: Some Statistical Methods and Techniques, edited by D. B. Owen
14. Sample Size Choice: Charts for Experiments with Linear Models, Robert E. Odeh and Martin Fox
15. Statistical Methods for Engineers and Scientists, Robert M. Betheoa, Benjamin S. Duran, and Thomas L. Boullion
16. Statistical Quality Control Methods, Irving W. Burr
18. Econometrics, Peter Schmidt
23. The Information in Contingency Tables, D. V. Gokhale and Solomon Kullback
25. Elementary Statistical Quality Control, Irving W. Burr
27. Basic Applied Statistics, B. L. Raktoe and J. J. Hubert
28. A Primer in Probability, Kathleen Subrahmaniam
30. Regression Methods: A Tool for Data Analysis, Rudolf J. Freund and Paul D. Minton
31. Randomization Tests, Eugene S. Edgington
32. Tables for Normal Tolerance Limits, Sampling Plans and Screening, Robert E. Odeh and D. B. Owen
33. Statistical Computing, William J. Kennedy, Jr., and James E. Gentle
34. Regression Analysis and Its Application: A Data-Oriented Approach, Richard F. Gunst and Robert L. Mason
35. Scientific Strategies to Save Your Life, I. D. J. Bross
36. Statistics in the Pharmaceutical Industry, edited by C. Ralph Buncher and Jia-Yeong Tsay
37. Sampling from a Finite Population, J. Hajek
38. Statistical Modeling Techniques, S. S. Shapiro and A. J. Gross
39. Statistical Theory and Inference in Research, T. A. Bancroft and C.-P. Han
40. Handbook of the Normal Distribution, Jagdish K. Patel and Campbell B. Read
41. Recent Advances in Regression Methods, Hrishikesh D. Vinod and Amanullah
42. Acceptance Sampling in Quality Control, Edward G. Schilling
43. The Randomized Clinical Trial and Therapeutic Decisions, edited by Niels Tystrup, John M Lachin, and Erik Juhl
44. Regression Analysis of Survival Data in Cancer Chemotherapy, Walter H. Carter, Jr., Galen L. Wampler, and Donald M. Stablein
45. A Course in Linear Models, Anant M. Kshirsagar
47. Statistical Analysis of DNA Sequence Data, edited by B. S. Weir
49. Attribute Sampling Plans, Tables of Tests and Confidence Limits for Proportions, Robert E. Odeh and D. B. Owen
51. Statistical Methods for Cancer Studies, edited by Richard G. Cornell
52. Practical Statistical Sampling for Auditors, Arthur J. Wilburn
53. Statistical Methods for Cancer Studies, edited by Edward J. Wegman and James G. Smith
55. Applied Factorial and Fractional Designs, Robert A. McLean and Virgil L. Anderson
56. Design of Experiments: Ranking and Selection, edited by Thomas J. Santner and Ajit C. Tamhane
58. Ensemble Modeling: Inference from Small-Scale Properties to Large-Scale Systems, Alan E. Gelfand and Crayton C. Walker
60. Bayesian Analysis of Linear Models, Lyle D. Broemeling
61. Methodological Issues for Health Care Surveys, Brenda Cox and Steven Cohen
63. Statpac: A Statistical Package for Microcomputers—PC-DOS Version for the IBM PC and Compatible, Bruce J. Chalmer and David G. Whitmore
64. Statpac: A Statistical Package for Microcomputers—Apple Version for the II, II+, and Ile, David G. Whitmore and Bruce J. Chalmer
66. Design and Analysis of Experiments, Roger G. Petersen
67. Statistical Methods for Pharmaceutical Research Planning, Sten W. Bergman and John C. Gittins
68. Goodness-of-Fit Techniques, edited by Ralph B. D'Agostino and Michael A. Stephens
69. Statistical Methods in Discrimination Litigation, edited by D. H. Kaye and Mikel Aickin
70. Truncated and Censored Samples from Normal Populations, Helmut Schneider
71. Robust Inference, M. L. Tiku, W. Y. Tan, and N. Balakrishnan
73. Assignment Methods in Combinatorial Data Analysis, Lawrence J. Hubert
74. Econometrics and Structural Change, Lyle D. Broemeling and Hiroki Tsurumi
75. Multivariate Interpretation of Clinical Laboratory Data, Adelin Albert and Eugene K. Harris
76. Statistical Tools for Simulation Practitioners, Jack P. C. Kleijnen
78. A Folio of Distributions: A Collection of Theoretical Quantile-Quantile Plots, Edward B. Fowkes
79. Applied Categorical Data Analysis, Daniel H. Freeman, Jr.
80. Seemingly Unrelated Regression Equations Models: Estimation and Inference, Virendra K. Srivastava and David E. A. Giles
81. Response Surfaces: Designs and Analyses, Andre I. Khuri and John A. Cornell
82. Nonlinear Parameter Estimation: An Integrated System in BASIC, John C. Nash and Mary Walker-Smith
84. Mixture Models: Inference and Applications to Clustering, Geoffrey J. McLachlan and Kaye E. Bastard
85. Randomized Response: Theory and Techniques, Aniruddha Chaudhuri and Rahul Mukerjee
87. Parts per Million Values for Estimating Quality Levels, Robert E. Odeh and D. B. Owen
89. Properties of Estimators for the Gamma Distribution, K. O. Bowman and L. R. Shenton
90. Spline Smoothing and Nonparametric Regression, Randall L. Eubank
91. Linear Least Squares Computations, R. W. Farnsworth
92. Exploring Statistics, Damaraju Raghavarao
93. Applied Time Series Analysis for Business and Economic Forecasting, Sufi M. Nazem
94. Bayesian Analysis of Time Series and Dynamic Models, edited by James C. Spall
95. The Inverse Gaussian Distribution: Theory, Methodology, and Applications, Raj S. Chhikara and J. Leroy Folks
96. Parameter Estimation in Reliability and Life Span Models, A. Clifford Cohen and Betty Jones Whitten
97. Pooled Cross-Sectional and Time Series Data Analysis, Terry D. Dielman
100. Nonlinear Lp-Norm Estimation, Rene Gonin and Arthur H. Money
101. Model Discrimination for Nonlinear Regression Models, Dale S. Borowiak
102. Applied Regression Analysis in Econometrics, Howard E. Doran
104. Statistical Methodology in the Pharmaceutical Sciences, Donald A. Berry
105. Engineering Design in Biotechnology, Perry C. Hasland
107. Handbook of Nonlinear Regression Models, David A. Ratkowsky
108. Robust Regression: Analysis and Applications, edited by Kenneth D. Lawrence and Jeffrey L. Arthur
109. Statistical Design and Analysis of Industrial Experiments, edited by Subir Ghosh
113. Engineering Quality by Design: Interpreting the Taguchi Approach, Thomas B. Barker
114. Survivorship Analysis for Clinical Studies, Eugene K. Harms and Adelin Albert
116. Stochastic Models of Carcinogenesis, Wai-Yun Tan
118. Handbook of Sequential Analysis, B. K. Ghosh and P. K. Son
119. Truncated and Censored Samples: Theory and Applications, A. Clifford Cohen
120. Survey Sampling Principles, E. K. Foreman
123. Handbook of the Logistic Distribution, edited by N. Balakrishnan
124. Fundamentals of Biostatistical Inference, Chap T. Le
125. Correspondence Analysis Handbook, J.-P. Benziéri
162. Probability and Statistical Inference, Nitis Mukhopadhyay
163. Handbook of Stochastic Analysis and Applications, edited by D. Kannan and V. Lakshmikantham
164. Testing for Normality, Henry C. Thode, Jr.
166. Visualizing Statistical Models and Concepts, R. W. Farebrother
169. Computer-Aided Econometrics, edited by David E. A. Giles

Additional Volumes in Preparation

The EM Algorithm and Related Statistical Models, edited by Michiko Watanabe and Kazunori Yamaguchi

Multivariate Statistical Analysis, Narayan C. Giri
Preface

Econometrics is an orderly fusion of economists’ mathematical models, data that measure economic activities of various types, and the tools of mathematical statistics. It is the empirical side of economic science. In the 21st century it is difficult to think of econometrics as being anything other than “computer-aided.” In fact, this description might seem almost redundant. Modern students and practitioners of econometrics find it difficult to conceive of life without almost unlimited, fast, and virtually free computing capacity. Moreover, we have come to expect this capacity to be right there on (or under) our desks. How could we practice our trade without the aid of a computer? Of course, as has been the case in most fields of endeavor in recent years, econometricians have simply expanded their horizons as computers have improved and become more readily accessible.

Econometrics is a relatively “young” discipline, dating from approximately 1930. During the 1940s and 1950s, for example, many new theoretical developments emerged in this fledgling field, ahead of the researchers’ ability to put them fully into practice in the context of real economic data. There were many obvious examples of this in the analysis of structural simultaneous equations models, for instance. There was a time when progress in econometrics was constrained by the available computing technology, at least in terms of the application of the theory. At one time, it was acceptable enough for an applied econometrician to comment that although (s)he was aware that a certain procedure should be used in a particular situation, this was not being done because of the unreasonable computational requirements or costs. Thankfully, today this is neither acceptable nor common. In the same vein, the application of conceptually simple statistical tools, such as Monte Carlo simulation, was
severely hampered by the high cost and low speed of computing. Consider, for example, the pathbreaking (and “capital intensive”) simulations of Summers* relating to the finite-sample properties of various estimators for simultaneous equations models, in which only 50 Monte Carlo replications were used. Compare this with MacKinnon’s response-surface simulations for unit root and cointegration tests, which involved a total of 15 million replications in 600 experiments for cases involving one to five time series, and a total of 12 million replications in 480 experiments in the case of six time series. And then think of the extent to which computing power has exploded in the last decade or so!

This book could, perhaps, have been titled \textit{Computer-Intensive Econometrics}, but this might have created the impression that the methods and applications that are presented in the various chapters that follow are ones that require access to exceptional computing facilities. For the most part, this is not so. It is true that there are many recent and ongoing major developments in econometrics, and in applied statistics generally, that take advantage of parallel processing and other “high end” aspects of computational science. Here, however, we are concerned with the use of readily accessible computing facilities. We are also concerned, though, with the creative use of these facilities and the development of algorithms and software that enable us to feel relatively unconstrained in our pursuit of new theoretical results and their application to economic data. Today, econometrics is “computer-aided” to a degree that even our middle-aged peers would never have dreamed possible.

This is a common story, of course, and there is undoubtedly much more to come. At this point in time, however, it is appropriate to reflect on the extent to which econometric techniques and computing methods and hardware are intertwined. Hopefully, the reader will find that the contributions in this volume offer some insights into this matter, and provide a wide range of important new results.

The contributors to this book represent a cross section of the international econometric community. In approaching potential participants I sought to include those who would bring a variety of views, backgrounds, and experiences to this venture. I am convinced that in this respect I have been successful, and I am extremely grateful to all of the contributors for their enthusiastic participation, for their patience, and for the fine chapters


that they have produced. In editing this volume I have benefited from the advice and assistance of a number of people. All of the chapters were anonymously peer-reviewed to ensure a high standard, and I would like to thank all of those who generously gave of their time and expertise in reviewing the draft chapters and providing feedback to me and to the authors. The editorial and production personnel at Marcel Dekker, Inc., were extremely encouraging, skillful, helpful and (above all) patient during the period that this volume was being prepared, and I thank them for making this book possible. I am especially grateful to Maria Allegra and Elissa Ryan at Marcel Dekker, Inc., and to Kelly Hall at Keyword Typesetting Services for all of their help.

David E. A. Giles
Contents

Preface iii
Contributors xi
Introduction xiii

1. Some Methodological Questions Arising from Large Data Sets 1
   Clive W.J. Granger

Applications of Simulation Methods

2. Finite-Sample Simulation-Based Tests in Seemingly Unrelated Regressions 11
   Jean-Marie Dufour and Lynda Khalaf

3. Finding Optimal Penalties for Model Selection in the Linear Regression Model 37
   Maxwell L. King and Gopal K. Bose

4. On Bootstrap Coverage Probability with Dependent Data 69
   Janis J. Zvingelis

vii
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norman R. Swanson, Ataman Ozyildirim, and Maria Pisu</td>
<td></td>
</tr>
<tr>
<td>Ron C. Mittelhammer and George G. Judge</td>
<td></td>
</tr>
<tr>
<td>7. Testing for Unit Roots in Semiannual Data</td>
<td>175</td>
</tr>
<tr>
<td>Sandra G. Feltham and David E. A. Giles</td>
<td></td>
</tr>
<tr>
<td>Bayesian and Related Inference</td>
<td></td>
</tr>
<tr>
<td>8. Using Simulation Methods for Bayesian Econometric Models</td>
<td>209</td>
</tr>
<tr>
<td>John Geweke, William McCausland, and John Stevens</td>
<td></td>
</tr>
<tr>
<td>9. Bayesian Inference in the Seemingly Unrelated Regressions Model</td>
<td>263</td>
</tr>
<tr>
<td>William E. Griffiths</td>
<td></td>
</tr>
<tr>
<td>Marco van Akkeren</td>
<td></td>
</tr>
<tr>
<td>Econometric Modeling</td>
<td></td>
</tr>
<tr>
<td>R. Glen Donaldson and Mark J. Kamstra</td>
<td></td>
</tr>
<tr>
<td>12. Neural Networks: An Econometric Tool</td>
<td>351</td>
</tr>
<tr>
<td>Johan F. Kaashoek and Herman K. van Dijk</td>
<td></td>
</tr>
</tbody>
</table>
Contents

13. Real-Time Forecasting with Vector Autoregressions: Spurious Drift, Structural Change, and Intercept Correction
   Ronald Bewley 385

14. Econometric Modeling Based on Pattern Recognition via the Fuzzy C-Means Clustering Algorithm
   David E. A. Giles and Robert Draeseke 407

Nonparametric and Semiparametric Inference

   Tae-Hwy Lee and Aman Ullah 451

16. The Effect of Economic Growth on Standard of Living: A Semiparametric Analysis
   Nilanjana Roy 479

Index 499
Contributors

Ronald Bewley, Ph.D. Professor, Department of Economics, University of New South Wales, Sydney, Australia

Gopal K. Bose, Ph.D. Statistician, Office of Economic and Statistical Research, Queensland Treasury, Brisbane, Queensland, Australia

R. Glen Donaldson, Ph.D. Finning Junior Professor of Finance, Faculty of Commerce & Business Administration, University of British Columbia, Vancouver, British Columbia, Canada


Jean-Marie Dufour, Ph.D. Professor, Département de Sciences Économiques, Université de Montréal, Montreal, Québec, Canada

Sandra G. Feltham, M.A. Senior Benefits Analyst, Ministry of Human Resources, Government of British Columbia, Victoria, British Columbia, Canada

John Geweke, Ph.D. Professor, Economics/Statistics and Harlan E. McGregor Chair in Economic Theory, Department of Economics, University of Iowa, Iowa City, Iowa, U.S.A.

David E. A. Giles, Ph.D. Professor, Department of Economics, University of Victoria, Victoria, British Columbia, Canada

Clive W. J. Granger, Ph.D. Professor, Department of Economics, University of California, San Diego, La Jolla, California, U.S.A.

William E. Griffiths, Ph.D. Professor, Department of Economics, University of Melbourne, Melbourne, Australia

George G. Judge, Ph.D. Professor, Department of Agricultural & Resource Economics, University of California at Berkeley, Berkeley, California, U.S.A.
Contributors

Johan F. Kaashoek, Ph.D. Professor, Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands

Mark J. Kamstra, Ph.D. Financial Economist and Associate Policy Advisor, Research Department, Federal Reserve Bank of Atlanta, Atlanta, Georgia, U.S.A.

Lynda Khalaf, Ph.D. Associate Professor, Département d’Économique, Université Laval, Quebec City, Québec, Canada

Maxwell L. King, Ph.D. Professor, Department of Econometrics & Business Statistics, and Deputy Dean, Faculty of Business & Economics, Monash University, Melbourne, Australia

Tae-Hwy Lee, Ph.D. Associate Professor, Department of Economics, University of California at Riverside, Riverside, California, U.S.A.

William McCAusland, Ph.D. Professor, Département de Sciences Économiques, Université de Montréal, Montréal, Québec, Canada

Ron C. Mittelhammer, Ph.D. Professor, Department of Agricultural & Resource Economics, Washington State University, Pullman, Washington, U.S.A.


Maria Pisu, Ph.D. Assistant Professor, Health Services Administration, University of Alabama, Birmingham, Alabama, U.S.A.

Nilanjana Roy, Ph.D. Assistant Professor, Department of Economics, University of Victoria, Victoria, British Columbia, Canada


Norman R. Swanson, Ph.D. Associate Professor, Department of Economics, Rutgers University, New Brunswick, New Jersey, U.S.A.

Aman Ullah, Ph.D. Professor, Department of Economics, University of California at Riverside, Riverside, California, U.S.A.

Marco van Akkeren, Ph.D. Post-Doctoral Fellow, Department of Agricultural & Resource Economics, University of California at Berkeley, Berkeley, California, U.S.A.

Herman K. van Dijk, Ph.D. Professor, Econometric Institute, Erasmus University Rotterdam, and Tinbergen Institute, Rotterdam, The Netherlands

Jānis J. Zvingelis, M.Sc. Ph.D. Student, Department of Finance, University of Iowa, Iowa City, Iowa, U.S.A.
Introduction

This volume comprises sixteen chapters, each of which demonstrates in one way or another the challenges of modern econometric practice, and the role that computer software and hardware play in this field. Without the extraordinary advances that have been made in computing in the last few years, econometrics would have been a very different discipline than the one it is today. Of course, there have been many significant developments in the application of mathematical and statistical tools to the theory of econometrics over this same time period. Moreover, there have been a number of crucial and quite dramatic shifts in the direction or emphasis of econometric methodology. With respect to the latter, some examples that come to mind include the emergence of the tools that we now take for granted in the area of non-stationary time-series analysis; the more central role that Bayesian methods have assumed within our discipline; and the growth in importance of methods for modeling with panel data.

Econometrics is “computer-aided” in many respects, and it is important to note that the relationship is by no means limited to empirical applications. On the contrary, as has been the case in many areas of statistical endeavor, computational methods and abstract theoretical developments often go hand in hand. For instance, the results of an exploratory Monte Carlo study may suggest that the sampling distribution of a new estimator or test statistic has certain desirable properties. This may be just the incentive that is needed for the exact theoretical results to be determined. This volume includes material that will illustrate this idea further. There are, of course many examples of developments in econometric theory that gained real significance only when computational techniques “caught up” to the extent that their implementation was feasible in realistic and relevant
situations. The estimation of simultaneous equations models using “full information” methods was an obvious example of this, as was the emergence of integration via importance sampling (and later, the Markov Chain Monte Carlo method) in the context of Bayesian econometrics.

In reality, there is very little activity in econometrics that is not “computer-aided” to some degree. However, this volume does not pretend to deal with the impact of computational tools in all areas of econometrics. This would be both pretentious and impractical. Instead, what is offered here is a relatively wide cross section of papers that reflect the role of computers and computational methods in modern econometric theory and applied econometrics. These papers vary considerably in their focus, but as a collection they should provide the interested reader with a clear impression of some of the more important ways in which computational issues are an integral part of econometrics.

The increasing availability of larger and larger sets of data has been both a blessing and a challenge for econometricians. Given that many econometric techniques are justified largely in terms of the desirable “large-sample” (asymptotic) properties that they possess, the limited size of the samples of data that most econometricians have had to work with historically has been problem-ridden and limiting. Indeed, this spawned a vast literature devoted to investigating the “small-sample” properties of these techniques. How relevant is all of this when data constraints begin to disappear? Not only have data sets become larger, but they have also changed in other ways. A good example of this is the availability of vast amounts of very high-frequency (e.g., hourly, by-the-minute, by-the-“tick”) data, especially in relation to financial markets. These data have features that do not arise when one has only a few dozen annual observations. This complication has to be balanced against the recognition that sometimes these very features enable one to construct models that are richer than would otherwise be possible, and to test them empirically.

In the opening chapter in this book, Granger looks critically at some of the issues surrounding the availability of very large sets of data for economic analysis. His comments are both timely and thought provoking. What does the availability of really large sets of data mean for the bulk of the material that is now taught in traditional econometrics courses? Does a large quantity of data necessarily mean that we will be able to test economic theories more effectively? How will we respond to the fact that with extremely large sample sizes, it is virtually impossible not to reject almost any hypothesis using classical methods? Granger reminds us that in this information age, with computing that is basically costless, and data sets that are beyond the wildest dreams that we may have had when we were in graduate school, we have to think afresh about many of the aspects of
Indeed, this sentiment is reflected in the rest of the material that the reader will find in this volume, which has been divided (albeit somewhat loosely, and with many overlaps) into four main parts:

1. Applications of Simulation Methods.
2. Bayesian and Related Inference.
3. Econometric Modeling.
4. Nonparametric and Semiparametric Inference.

The first of these parts includes six chapters, and this in itself speaks to the extent of the impact that simulation techniques have had on econometrics in recent years. The contributions here are quite varied in terms of specifics, but in each case there is a clear reliance on simulation to determine new and important results. Monte Carlo simulation, in particular, receives a considerable amount of attention in this part of the book. In Chapter 2, Dufour and Khalaf consider the problem of constructing tests of quite general hypotheses about the parameters of the well-known and widely used “Seemingly Unrelated Regressions” (SUR) model, first proposed by Zellner (1962). They are especially concerned with the performance of such tests in finite samples, and Monte Carlo methods are used in two ways in their analysis. Exploiting their earlier research into such problems in the context of the multivariate regression model (e.g., Dufour and Khalaf, 1998), they first develop a bounds test that is based on the likelihood ratio criterion. The need for a bounds test, in general, arises from the fact that the exact null distribution for the test statistic involves “nuisance parameters.”

The distribution of the bounds statistic itself is readily simulated by the Monte Carlo method, thus providing a rather nice example of the blending of formal statistical theory and computer simulation in order to obtain a strong and useful result. Of course, any “bounds test” will not always yield a result – the test statistic may lie in the “inconclusive region.” To deal with this possibility, Dufour and Khalaf again resort to Monte Carlo methods to get exact results, one of which effectively involves obtaining the associated $p$-value for each point in the nuisance parameter space, and then taking the maximum of these $p$-values. The techniques that the authors describe in this chapter clearly have widespread application to other complex testing problems in econometrics, and they provide some insights into the type of computer-aided testing that we are likely to see much more of in the future.

Model selection, rather than hypothesis testing, is the subject of Chapter 3 by King and Bose. One of the most common problems in applied econometrics, and in regression analysis generally, is the choice of which model to use. More specifically, we are often faced with a number of
plausible competing models, but we need to choose one of them as the basis for further analysis. When the competing models are “non-nested,” or when their structure precludes the use of standard hypothesis testing for other reasons (e.g., in the selection of the number of “augmentation” terms in a Dickey-Fuller regression), econometricians often turn to “Information Criteria” (IC) as the basis for model selection. Various IC measures are commonly in use, such as those associated with Akaike (1974), Schwarz (1978), Hannan and Quinn (and the Bayesian IC (BIC)). These criteria are likelihood-based, with different penalty functions for the model’s lack of parsimony, and for the most part they have only large-sample justification. King and Bose propose the use of simulation methods first to determine the probabilities of correct model selection in finite samples, and second to derive penalty values for IC procedures that are optimal in the sense of maximizing these probabilities on average. Simulated annealing is used in the latter computations, and a Monte Carlo study is conducted to determine the relative merits of the procedures. This study suggests that the new “optimized” procedure proposed by the authors can work very well if the sample size is quite small, but that the BIC is close to being optimal in about two thirds of the situations investigated.

The bootstrap resampling procedure proposed by Efron (1979) has been a key development in statistics, and it has found widespread application in many areas of econometrics. In its basic form, the bootstrap relies on the use of a random sample of data, but it can be modified in various ways if this assumption is not satisfied (as is often the case with economic data). In Chapter 4, Zhang takes up this point in the context of the use of the bootstrap to obtain finite-sample critical values for tests (or, equivalently, finite-sample coverage probabilities for confidence intervals). The question that he addresses is whether or not the known advantage of the bootstrap over the use of asymptotic (large n) approximations, in the random sample case, carries over to situations involving dependent data. He uses the non-overlapping block bootstrap to deal with the dependency issue, and uses Edgeworth expansions to prove that in this case the bootstrap provides only a small improvement over the coverage probabilities afforded by the asymptotic approximation. More specifically, he shows that while the errors associated with one-sided and two-sided asymptotic confidence intervals are reduced from $O(n^{-1/2})$ and $O(n^{-1})$, respectively, to $O(n^{-2})$ for a wide class of problems in the random sample case, they are reduced only to $O(n^{-3/4})$ and $O(n^{-4/3})$ in the context of dependent data. This has practical implications for the use of the bootstrap in this role in many areas of econometrics.

In Chapter 5, Swanson, Ozyildirim, and Pisu offer a very extensive set of Monte Carlo results that provides evidence relating to several questions that arise when a vector autoregressive (VAR) model is used as
the basis for testing for Granger (non-) causality. Such tests have attracted a great deal of attention in the econometrics literature, and elsewhere, and their implementation has had to be reconsidered in the context of data exhibiting unit roots and/or cointegration. One of the issues that has to be addressed is how to determine the lag lengths for the VAR model that serves as the vehicle for the causality testing. Another issue is how one should deal with the fact that the Wald test, that is usually used when testing for Granger noncausality, does not have its usual distributional properties (even asymptotically) if the data are nonstationary. With regard to the latter matter, the authors’ results provide very strong support for the simple modification to the testing procedure that was proposed by Toda and Yamamoto (1995). This support holds whether or not the non-stationary data are cointegrated. They also find that the Schwarz (1979) information criterion (SIC) performs well as a basis for testing Granger causality. Specifically, they favor using the SIC to find the preferred VAR model, and then rejecting the null hypothesis of noncausality if the supposedly causal variable appears appropriately in that model. Swanson et al. provide a large number of results that add significantly to our knowledge of this important topic. In doing so they also remind us that Granger’s (1969) original discussion of his concept of causality focused on out-of-sample predictive issues, and they incorporate this important feature of Granger causality into their own study.

Chapter 6, by Mittelhammer and Judge, will no doubt introduce many readers to a type of inference that they have not previously encountered in the econometrics literature. “Empirical Likelihood” (EL) methods of estimation and testing have been pioneered by Owen (1991, 2000) and a handful of other statisticians, but only very recently have these developments been recognized by econometricians. (See Mittelhammer et al., 2000, for an excellent discussion.) Essentially, EL methods lie part way between conventional likelihood methods (such as maximum likelihood estimation, and the likelihood ratio test) and nonparametric methods. One can argue that while the former methods have desirable asymptotic properties, these generally apply only if the underlying data-generating process is specified correctly. That is, if the form of the likelihood function is known. On the other hand, one can avoid the need to specify a parametric likelihood function by adopting nonparametric methods (such as kernel regression), and still retain some weak large-sample results. EL represents a compromise between these two polar positions, and the hope is that one will retain the best of each, rather than the worst of each approach! In general, this hope can be shown to be justifiable. Mittelhammer and Judge apply the EL approach to simultaneous equations models, and compare the small-sample properties of the maximum EL estimator with those of Two Stage Least
Squares (TSLS). EL methods involve computational issues beyond those associated with traditional estimators such as TSLS, so this in itself makes this study particularly relevant for inclusion in this volume. Moreover, the authors use a carefully designed Monte Carlo experiment to simulate the estimator properties—hence the inclusion of this chapter in this part of the book. Mittelhammer and Judge find that the maximum EL estimator fares well in comparison with the TSLS estimator, and they leave us with some interesting avenues for further research in this newly emerging area of econometrics.

As was the case in Chapter 5, nonstationary time-series data are also the focus of the last chapter in the first part of this book. In Chapter 7, Feltham and Giles use both analytical methods and Monte Carlo simulation to investigate the properties of some tests for various types of nonstationarity that can arise with time-series data that are recorded semiannually. Specifically, they look at individual and joint tests for unit roots at the zero and seasonal frequencies, determine their (nonstandard) asymptotic null distributions in terms of functionals of standard Brownian motions, and derive the associated critical values for a range of sample sizes and situations. The latter are obtained from a large Monte Carlo experiment, and the same approach is used to evaluate the power functions of these tests under various scenarios. The results that the authors obtain are linked to other results in the literature on stochastic seasonality, and some simple illustrations using actual data are also provided.

The contributions in the second part of this book are devoted to Bayesian (and related) methods, and these are dealt with in three chapters. The use of the tools of Bayesian statistical inference in the econometric setting dates from the late 1960s, and was energetically promoted from the outset by Zellner (1971) and his students and colleagues. Today, it is widely accepted that the gap that once separated Bayesian econometricians from the majority of their colleagues has all but vanished. I can attest to this with some personal knowledge, having completed my own doctorate in precisely this area of econometrics in 1975! An analogy might be drawn here with another major development in econometrics during roughly the same time period. Thirty or forty years ago there was something of a schism between econometricians who espoused the use of “pure time-series” methods (especially for forecasting purposes), and those who stressed the essential need for complex structural models. When we look at the way in which econometrics has developed, especially since the beginning of the “unit root/cointegration revolution,” we see that there has been a healthy blurring of this earlier artificial boundary between these two schools of thought, as can be seen in many of the contributions in this volume. Much the same can be said of the division between Bayesians and
“non-Bayesians” in econometrics, and elsewhere. Econometricians are now often quite willing to “change hats” as the situation demands. It is accepted that there are benefits in having access to a range of different tools, so that one can apply the one best suited to a particular task. Even our econometric language is immensely more inclusive, in this respect, than used to be the case.

As was noted in the Preface to this volume, the extensive integration of Bayesian inference into econometrics is a wonderful example of the positive impact that enhanced computing facilities have had on the discipline. Having been developed to a certain stage, Bayesian methods became somewhat difficult to “sell” to the rest of the profession (especially applied researchers) because of the computational difficulties associated with applying them to realistically complex models. Put simply, unless one was prepared to severely limit the way in which one’s prior information was expressed, and introduced into the analysis, one was limited to dealing with models with a tiny number of unknown parameters. The introduction of Monte Carlo integration, and later the revolution associated with Markov Chain Monte Carlo (MCMC) methods, changed all of this. Today, we regularly see fully-fledged Bayesian analyses of extremely complex models. Students in econometrics have ready access to these techniques, just one example being through the inclusion of Geweke’s (1989) Bayesian regression estimator in the SHAZAM (2001) computer package.

With these remarks in mind, it is most appropriate that this section of the book begins with Chapter 8, by Geweke, McCausland, and Stevens. This chapter begins by providing a clear overview of the basic components of Bayesian inference, and then discusses in some detail the simulation techniques (such as acceptance and importance sampling, and examples of MCMC methods) noted above that have revolutionized this area of econometrics. The use of Geweke’s Bayesian econometrics software (the BACC package) is illustrated through some interesting applications, and the reader will be left with a clear impression of the power and flexibility of this software with regard to both estimation and model selection. One could argue that this chapter represents computer-aided econometrics at its best, and those readers who have little inclination in the Bayesian direction will find their lack of faith severely tested!

In Chapter 9, Griffiths provides material that complements the previous chapter extremely well. His concern is with Zellner’s (1962) SUR model, discussed above in the context of Chapter 2. As he says, his objective is “…to provide a practical guide to computer-aided Bayesian inference for a variety of problems that arise in applications of the SUR model.” Griffiths provides us with insights into some of the practical issues that have to be faced when applying these Bayesian methods, including those that arise
when extending the analysis to models that are nonlinear. Three applications serve their purpose admirably, and will leave the reader with a good appreciation of the merits of modern Bayesian econometrics.

One of the key features of the Bayesian methodology is the flexible way in which prior information is “injected” into the analysis, both through prior masses for competing models, and through prior densities to represent the uncertain information about the parameters of these models. This prior information is then combined with the data information, using Bayes’ Theorem, and the resulting posterior distributions provide information that may be used (in conjunction with a chosen loss function) to draw inferences about parameters and models. If there is insufficient information in the sample to allow the estimation of all of the individual parameters in a model of interest, then one way to deal with this is to introduce a sufficiently informative prior density, and construct the Bayesian posterior distribution. A case in point is the use of the natural-conjugate prior in the context of a standard regression model whose regressor matrix is rank-deficient. Perfect “multicollinearity” is nothing more than a shortage of information, relative to the proposed task at hand. Often, such “ill-posed” problems can be solved in this way, but there are also alternative ways to proceed, as van Akkeren explores in Chapter 10. Here, he discusses the Data Based Information Theoretic (DBIT) estimator, that has been shown elsewhere to perform well (even in small samples) in the context of ill-conditioned linear inverse problems. This estimator uses the Kullback-Leibler information measure (e.g., Kullback, 1959) in its objective function, and is a competitor not only of the Bayesian approach, but also other well known techniques such as ridge regression (and shrinkage estimators in general), and principal components regression. Apart from providing an interesting and timely discussion of this important, and relatively new, approach to inference, van Akkeren also illustrates its application to two well known sets of data. One of these is also used by Giles and Draeseke in Chapter 14 of this volume, and the reader may wish to compare these two sets of results. The emergence of the DBIT estimator, and other important related developments in the area of inverse problems, would not have been possible without our modern computing software and hardware. Again, this chapter clearly illustrates yet another direction in which modern econometrics has been “computer-aided.”

The third part of this book deals with various modeling techniques that can be applied to economic data, and the four associated chapters include a number of empirical applications as well as proposing some new techniques. One of the areas where econometric techniques have found considerable application in recent times is finance. “Financial Econometrics” is a discernible subfield in its own right, and in Chapter 11
Donaldson and Kamstra provide a very interesting example of work in this area—one that is computationally intensive and creative. They are concerned with an important question that arises in financial economics, namely “are the market prices of stocks in line with the underlying ‘fundamentals,’ or are stock markets excessively volatile, perhaps containing price ‘bubbles’?” Readers who have followed or participated in equity trading in the past five years will need no convincing that this is an interesting question! The methodology that Donaldson and Kamstra adopt is intriguing. They use actual financial data to estimate time-series models for dividend growth and discount rates. These are used to simulate paths for dividend growth and discount rates for various “bubble-free” economies, and fair-value prices for equities in these simulated economies can then be determined. Essentially, they then compare market prices (for S&P 500 stocks) with simulated fundamental prices and look for any discrepancies. Their main findings are that market prices are actually not excessively volatile, and that the tests that other authors have used tend to be biased in the direction “finding” non-existent bubbles and excess volatility. The extent to which this type of analysis is “computer-aided” becomes apparent when one considers that Donaldson and Kamstra simulated approximately one million hypothetical economies, an exercise that involved a month of CPU-time on a SUN UltraSparc workstation.

The remaining three chapters in this part of the book deal with three different approaches to econometric modeling: the use of (artificial) neural networks to model complex nonlinear systems; the use of VAR models for real-time forecasting; and a new technique that adapts ideas from fuzzy set theory and the pattern recognition literature to help estimate nonlinear economic relationships. The first of these topics is covered in Chapter 12, by Kaashoek and van Dijk. Readers who are not familiar with the use of neural network models in econometrics will welcome the very accessible overview that these authors provide in the early parts of this chapter. They then move on to illustrate how neural networks can be used successfully to deal with the complex patterns that are associated with economic data. In particular, Kaashoek and van Dijk show how this type of analysis can be used to uncover the dynamic properties of highly nonlinear systems, and they provide some interesting empirical applications relating to exchange rate data and a “Phillips curve” for the U.S.A.

In Chapter 13, Bewley discusses some important issues that arise when using a VAR model for forecasting purposes. This chapter covers a very wide range of issues—even wider than is suggested by its title. Among the topics that the reader will encounter here are Bayesian VAR models, nonstationary time-series data, structural breaks, and macroeconomic forecasting. Bewley pays careful attention to the role of drift (intercept)
terms in the equations of a VAR model, and proposes an automated procedure for ensuring that structural changes are taken into account in a manner that will enhance forecast performance. The merits of this proposal are supported by the results of a Monte Carlo experiment. Notably, Bewley finds that adjustments to the intercept terms of a VAR model estimated in the differences of the data, with the intention of bringing the model “back on track,” lead to an improvement in forecast mean squared error. Previous work by Clements and Hendry (1996) had suggested that this would be a useful procedure only for VAR models specified in the levels of the data. There are aspects of this chapter that will help younger readers appreciate some of the historical connections in the econometrics literature, and econometric practice. By way of example, we are reminded that post-estimation intercept corrections (based on the most recent forecast residuals) were standard practice in the use of the large-scale structural forecasting models of the late 1960s and 1970s. I recall this well from my own experience in the development and use of the Reserve Bank of New Zealand’s econometric model at that time. Bewley points also out that the popular Bayesian VAR modeling procedure is really nothing more than a special application of the Theil and Goldberger (1961) “mixed regression” method, using appropriate dummy variables. This is not to say that there is “nothing new under the sun”, but it is helpful to have an historical perspective on new developments in any discipline.

Chapter 14, by Giles and Draeseke, completes the third part of this volume. As noted above, it considers some new ways of modeling non-linear economic relationships, by drawing on various concepts from the literature relating to fuzzy sets and fuzzy logic, and that associated with pattern recognition and image processing. The authors show how a fuzzy clustering algorithm can be used to partition a set of data in a flexible manner, and then fuzzy logic can be used to combine separate models estimated from each cluster into one overall model. Because the combining process involves weights that vary continuously through the sample, highly nonlinear characteristics can be captured very successfully, even if the individual component models are themselves linear. Giles and Draeseke illustrate this technique with a range of empirical applications, including simple demand for money and consumption models, trend fitting, and an analysis of the age-earnings data that is also analyzed by van Akkeren in Chapter 10. They find that their “fuzzy regression” approach outperforms nonparametric kernel regression in terms of “goodness of fit.” Perhaps even more importantly, they show that this new technique can be applied to multivariate relationships that are beyond the scope of standard nonparametric regression (due to the well-known “curse of dimensionality” associated with the latter).
The final part of the book includes two chapters that reflect the important impact that nonparametric methods have had on econometric theory and practice in recent times, and also fit together remarkably well as a pair. Given that there are substantial penalties (e.g., bias, inconsistency) that are usually incurred if we misspecify the functional form of an econometric model or omit relevant regressors, an attractive alternative may be to use nonparametric kernel regression. In Chapter 15, Lee and Ullah consider various ways of using nonparametric regression not just as an alternative to parametric regression in its own right, but also as a basis for testing for different forms of model misspecification. This testing can be undertaken by using one or more sets of nonparametric regression residual vectors or prediction vectors. Clearly, classical (parametric) hypothesis testing cannot be used, and the authors investigate three different ways of dealing with this problem. In each case, bootstrap methods are used to implement the tests—specifically to determine p-values for finite-sample situations. Lee and Ullah conduct a sizeable Monte Carlo experiment to determine the relative effectiveness of each of these three testing procedures, and they find that the class of tests nonparametric tests proposed by Li and Wang (1998) and by Zheng (1996) generally dominates the other two that are considered. This study provides an insightful “bridge” between parametric and nonparametric regression with respect to an important aspect of econometric analysis. It also illustrates once more the critical role of computational/simulation techniques in modern econometrics.

This volume concludes with an empirical study that (fortuitously, perhaps) makes particular use of the bootstrap specification test of Li and Wang (1998) that emerges so well from the study in the previous chapter. In Chapter 16, Roy returns to a long-standing question from the economics literature: “Does an increase in per capita income lead to an improvement in standard of living?” In this context, “standard of living” has been measured in terms of various individual indicators (such as life expectancy at birth), or composite indices of indicators, by different authors. One of the issues associated with the previous literature on this topic has been the choice of functional form when measuring such a relationship, and with this in mind the use of conventional nonparametric regression has been pursued by other researchers. In this chapter, Roy tests a linear parametric model against a nonparametric specification, using the testing procedure of Li and Wang. She also considers a semiparametric model that includes both per capita income and per capita public health expenditure, the latter entering non-linearly. This study underscores the importance of undertaking appropriate specification testing, rather than simply abandoning a parametric approach and using standard nonparametric regression. The nonparametric approach that Roy takes also has the merit of at least partly dealing with one of the
major limitations of kernel regression—the so-called “curse of dimensionality” that was mentioned above in connection with the research of Giles and Draeseke, reported in Chapter 14.

In many ways it is fitting that this volume ends with this empirical study. Roy’s use of an interesting (though rather short) panel of data illustrates some of the data issues that need to be addressed properly in modern econometric applications. Our present level of theory, and the readily available computing facilities, allow one to be flexible as to the choice of technique that is adopted—standard methods can be reconsidered, adapted and applied with comparative ease. Econometrics as a discipline is truly “computer-aided,” and the practicing econometrician is freed from many of the constraints that previously limited our horizons. With this freedom comes the opportunity for enhanced and responsible creativity as we push the boundaries of empirical economics.

REFERENCES

1

Some Methodological Questions Arising from Large Data Sets

Clive W.J. Granger
University of California, San Diego, La Jolla, California, U.S.A.

1 INTRODUCTION

In recent years there have been strong trends in both data availability and computing speed, which will have had dramatic effects on how econometricians approach their tasks. These trends are very likely to continue. To make my discussion easier I will assume that, in the future, computing will be sufficient for all our needs and will be effectively free. These features are almost true now. My home university has a 1 teraflop computer (meaning that it can perform one trillion calculations per second) available for academic use and U.S. government research agencies have a 4.5 teraflop machine available now with a plan for a 100 teraflop machine to come on line in 2004. IBM has already announced work on Blue Gene, claimed to be 500 times faster than any present machine.

Suppose that a statistician or econometrician is faced with the prospect of having enormous quantities of data combined with effectively unlimited computing abilities, what reaction should one expect? Initially there would be simple delight, as one realizes that many of the present constraints on analysis have been removed. However, on reflection, the reaction could be
more muted when the fact that it is unclear what techniques should be used on the data. It is also very unclear if the methods traditionally taught to and by econometricians will have any relevance. In the past, the major occupation of econometric theorists was in developing estimation techniques for many different situations and satisfactory tests to be used for inference. As has been pointed out several times before, as in Granger (1998), for a large enough data set it is virtually impossible to not reject a specific null hypothesis. For example, consider a credit card company that observes one million transactions per day and is interested in the proportion of transactions above a certain monetary amount, say $200. Suppose that recently this amount has been 41.4% and that yesterday the observed proportion was 0.4152. Denoting this proportion by \( \hat{p} \), using the normal approximation for the binomial, the 99.9% confidence interval is \( \hat{p} \pm 3.30 \sqrt{\hat{p}(1-\hat{p})/n} \) for a sample size \( n \), with \( \hat{p} = 0.4152 \) and \( n = \) one million; this gives a band of \( 0.4152 \pm 0.001 \), which does not include 0.414. Thus, to a high level of statistical significance, there appears to have been a change in the proportions.

This example, and similar thoughts, lead to the following conclusions that have been discussed in Granger (1998):

1. Limit theorems and asymptotic theory generally will be appropriate with very large samples. See Lehmann (1999) for a recent discussion. Conversely, all the small-sample procedures, which take up so much time in elementary statistics courses, are irrelevant.
2. One can expect that virtually any explicit hypothesis will be rejected, such as \( H_0: \text{mean} = 3 \), but a broader null \( H_0: \text{mean in range} (2.5–3.5) \) may not be.
3. Statistical significance becomes less relevant compared to economic significance, even though the latter is difficult to quantify in most circumstances.
4. Any consistent estimate will provide satisfactory values; efficiency estimation is not particularly important.
5. It is better to use estimates, and model specifications, that depend on fewer, or looser, assumptions.
6. Bayesian priors become irrelevant unless the number of parameters to estimate is infinite, as in a moving average (MA) (\( \infty \)) model.
7. The “pooling” assumption of many standard panel methods will be rejected, requiring a joint analysis of many time series from different “agents.”

As inference, testing, and estimation have been the life-blood of econometricians for at least the past 50 years, to be reminded that they are
Questions Arising from Large Data Sets

of little importance with large data sets may seem to be a major turn-off. There are a number of reactions to this situation; it can be regarded as an exciting opportunity to consider new types of specifications, methods, models, and situations some of which will be considered below. Alternatively, one can ask if something like the old methodology and approaches can be retained and used under particular circumstances.

The answer is obviously yes because however large an information set is, there will usually be sensible ways to subdivide it into subsets that are small enough for classical techniques to be appropriate. Examples would be:

1. A large cross-section of marketing questions involving one million individuals could be divided into many geographical regions, subdivided into income groups, genders, racial, or country of origin groups, and so forth. Eventually, the average number in each subgroup may be fairly small, just a few hundred. What is less clear is if anyone really cares about the average juice consumption of right-handed male Californians originally from Britain with an income in the top 10%, for example, compared to the consumption of some other group. No doubt statistical significances could be found and regressions run but the economic or marketing significance may be slight.

2. A large panel of multinational companies could record monthly sales, changes in stock prices, quantity of shares traded, CAPM beta values, and so forth, over 20 years. The panel members could be considered in various subgroups, depending on size, work force, and country of the home office, and for different subsets of years.

3. A very long single time series, such as daily volume of trading on some exchange, or the percentage of shares increasing over those decreasing, could be recorded from 1927 to 2000, for example, giving over 18,000 daily values. If one uses a running window of, say, 1000 terms and within each window applies standard modeling techniques, a time-varying parameter time series model will be achieved. This model can be tested for significance of changes and interpretations, hopefully of economic relevance, given to the swings in coefficient values or structural changes.

These examples also illustrate a problem with the definition of a large data set. In economic time series, a size of 1000 would be considered large, in fact virtually impossible to achieve in macroeconomics. Only recently in finance have the truly huge data sets become available. In cross-section econometrics, sets involving 10,000 individuals are not unusual, and if one is working with census data, the sample could be in the tens of millions. In this area, large samples are not a novel problem. In panels, the cross-sectional
dimension is often quite large, but the time series dimension rather small, though new panels will be large in both dimensions. This area will present the greatest challenges.

To preserve the usefulness of their present human capital, econometricians may concentrate on techniques of great interest but with very slow convergence rates, such as some nonparametric problems and questions involving extreme values, which can have convergence rates of $n^{1/3}$ or less.

Consideration of such possibilities are potentially important in the short-run, but cannot hold back the inevitable development of new techniques.

2 TWO TIME SERIES SCENARIOS

Two hypothetical experiments will be considered:

1. Starting at a fixed point in time, we observe a vector series $X_t$ at a fixed observation interval, say monthly, giving a sample length $n$. Initially, $n$ is quite small, say 50, but after a period it becomes, 200, then 500, and eventually the series consists of 2000 months, say, or possibly more. The question to consider is: what could a time series analyst do differently when $n$ is larger than when it was smaller?

2. The second mind experiment takes a set period of time $(0, T)$ and initially observes it annually, say, giving $X_t$, $t = 1, \ldots, T$. We then observe it monthly giving $12T$ pieces of data (roughly), then weekly giving $52T$ points, then daily, so that there are now $365T$ points (or $216T$ if we just use weekdays), roughly $2000T$ hours, and so forth.

In each experiment we are getting more numerical information but of a different nature.

For the first experiment, the types of models and techniques used can increase in complexity and sophistication as the sample size increases. For very short series, a simple autoregressive model of low order plus trend is appropriate; for a somewhat larger series, seasonal components, unit root testing, and autoregressive integrated moving average (ARIMA) models come into play. When the series is somewhat larger, one has predictive conditional distributions, quantile estimates, large vector autoregression (VAR) models, complicated nonlinear models, time-varying parameter models, breaks, regime shifts, outliers, and so forth. However, although it is clear that the variety of models available for consideration will increase with sample size, the above discussion is in terms of the viewpoint of an
idealized statistician as it is context-free. In practice, as the data set gets larger, one has to be concerned with the possibility that the economy that produced the early terms in the series is quite different from the economy described by the later terms. Economies evolve as the tastes of its agents and the form of its institutions change. A form of time-changing parameter model may capture this, but this essentially shortens the series, as mentioned in Section 1.

In the second experiment, as the observation period is reduced, one obtains more information about the high-frequency part of the spectrum, but no more about the low-frequency components. For example, there is no information about the weekly cycle in data observed at a period of a week or more, but daily data will contain a lot. In a sense, this example illustrates the idea that one can have more data but no more information. If one is looking for exceptionally large rises or drops in stock prices, for example, you do not find more of them by looking at a 5 minute price index rather than an hourly index, and possibly a daily index. The extreme case is the use of tick-by-tick data, called “ultra-high frequency data” by Engle (2000). These are unusual data because it is not clear if the information about the \( j \)th price of the day was available to the decision maker(s) who produced the trade for the \( j+1 \)th price, because of the speed required and as there can a queue of as yet unreported trades. That is, the time to decide to make the trade, the time of the trade, and the time the trades reported are by no means kept in the same order. There is certainly a huge amount of tick-by-tick data; its value-added is still being determined. It seems to be clear increasing the frequency of data observation is helpful to some aspects of data analysis, but not others. As an analogy, one could ask if daily data are more useful for studying business cycles than monthly data, except for the fact they are probably issued with a shorter delay.

3 ONWARD AND UPWARD

To reiterate, or expand on, some earlier points, having a lot of data is an opportunity, not a burden. Given the choice of doing the old things better or attempting to do something new, I hope we will chose the latter. Some of the things that we can do with large data sets that are not possible with fewer data include:

1. Consider a wider variety of specifications.
2. Discover more subtle effects.
3. Distinguish between close specifications.
4. Obtain better estimates of parameters.
5. Do more and obtain more useful post- and out-of sample evaluations.
6. Attempt types of analysis that are now not possible.

Details in some cases will be discussed below. There are also a number of general observations that could be made frequently but will just be recorded here:

1. As stated before, standard statistical tests are no help with very large sample sizes, as any simple null hypothesis will almost certainly be rejected. If the null is taken to be the simple case — normality, linearity, stationarity, few explanatory variables — e.g., if these are all rejected one ends up with complicated situations, particularly nonlinear, nonstationary models with many related variables.
2. For panel data, with both many \( N \) agents and long time \( T \) dimensions, it follows from (1) that the poolability assumption will be rejected, so that standard panel analysis techniques are largely irrelevant. The data will have to be treated like a vector of time series, presumably inter-related.
3. As there will be many model specifications available, a precise statement about the purpose of the analysis is essential before starting. The purpose is also necessary for the evaluation phase of the modeling process.
4. Evaluation should be in terms of economic significance of alternative models rather than statistical significance. Economic significance can, theoretically at least, be judged in terms of the economic values to decision makers who base their decisions on the outcomes of the models. This perspective is easy in finance and forecasting, but needs further development in other areas of econometrics.
5. As model specifications may well be complicated, presentation of results from these may be difficult. Innovations in presentations of parametric values or outputs of alternative complex models will be needed. The use of mobile trees that increase in complexity about information in a region as one moves around the branches are already successful in containing a great deal of facts in a readable form. Animation of diagrams may also be helpful.
6. The quality of data will not necessarily change as data sets get larger, but running filters to discover missing values, misprints or other data errors, outliers (however defined) will be more difficult to generate in larger multidimensional sets and arbitrary fixes may
be used to remove them. However, using outlier-resistant methods will not affect the efficiency of estimates as there is too much information, unless one is specifically modeling outliers, such as in value-at-risk.

7. With a fairly small data set, the analyst becomes very familiar with its properties and may “find” something unexpected, which occasionally continues to occur out-of-sample. Serendipity has been an important tool in applied statistics but will almost certainly be lost with the advent of very large data sets. Computers can be programmed to search for unexpected events but in a sense they have to be in some way expected to be considered for a search.

8. The answer to the question “Are there too many data?” will depend on the purpose of the analysis and the needs of the decision makers who will use the output. If the purpose is vague, such as “help understanding,” the question cannot be answered. For many purposes, having more data does not imply that we have more relevant information, we may just be adding noise. For other purposes, precision may be important and this can increase with sample size.

4 CROSS-SECTION DATA

This area of econometrics already has experience with large data sets. Suppose that we have a vector of $m$ dependent variables $Y$ and a further vector of $k$ explanatory variables $X$. The objective may be thought of as making statements about the conditional distribution of $Y$ given $X$. The question is not often tackled in terms of conditional distributions, but rather in conditional moments, such as means, variances, and covariances. The number of specifications to model each is still very high and usually limiting classes are considered, such as linear or quadratic terms of the explanatory variables, based possibly on an assumption that the conditional distribution is Gaussian, even though such an assumption is likely to be rejected by the data. The aim in all approaches is to consider a wide enough class of specifications so that, hopefully, a reasonable approximation to the actual data-generating process will lie with this set. Given enough data it is hoped that one can sort through this collection of models, discard the unsatisfactory ones, and end with one that is suitable for use. The sorting may use the “general to simple” methodology, which has strong theoretical and practical foundations in the class of linear models, but is more difficult to use with nonlinear models, as the number of alternatives is much greater and
there are less clear hierarchies. It should be noted that with large data sets there are fewer of the classical data-mining problems, with an inferior model appearing to be the best. Instead, one may obtain several specifications that appear to be different but which have similar statistical properties, and a choice between them in terms of economic benefits may be difficult. The spurious results in small samples are replaced by indifference between specifications when using large samples.

5 TIME SERIES DATA

For time series, methodological problems arise both from the length of the series, as discussed above, and also from the number of series involved. The problems can be illustrated using a simple linear VAR model with a vector of \( N \) series. If a full model, with constant parameters, is constructed with \( p \) lags for each variable in each equation, the total number of parameters is \( N^2p \) — known as the curse of dimensionality, even without considering estimation of the covariance matrix of the residuals. The value of \( p \) is not really a choice; it is the value required to whiten the residuals. For example, if \( N = 2000, p = 10, N^2p = 4 \times 10^7 \). I chose this example because there are roughly 2000 shares quoted in the New York Stock Exchange. A VAR for absolute daily returns would need \( p \) larger than 10. The problems that arise concern estimation, presentation, and interpretation of so many parameters plus how to compare alternative specifications. The solution used in building large economic models is to be parsimonious with the inclusion of variables in equations. One might always include the lagged dependent variables, lags of closely associated variables (e.g., from economic theory, or close trading partners for countries, or in some industry for companies), and, most helpfully, a few “common factors” which efficiently mop up inter-relationship between many variables, such as the CAPM model and its extensions. Using such a strategy, which can either be “hands on” or automated, the number of parameters estimated can be reduced to \( mNp \), where \( m \) is some integer in the range 5–10, giving a reduction of parameters of over 99% in the example. The economic loss is unclear.

The linear, constant parameter model for the conditional mean, as just considered, is almost certainly not where the future of time series lies. With lots of data, subtle nonlinearities and/or time-varying parameters can be detected (sometimes they may be alternative specifications) and complicated models specified and fitted, probably recursively. Concentration on models for the conditional mean and the conditional variance will evolve into attention being given to other moments and properties of the conditional or predictive distribution, particularly the quantiles. An approximation to the
whole predictive distribution will result, not necessarily with an assumption of normality, and with time-varying parameters. The users will concentrate on the quality and usefulness of the output rather than on the specification of the model, which will be very complicated. Alternative models will be compared by the relative usefulness of their outputs for decision making; purely statistical measures are unlikely to have plausible economic interpretations.

6 CONCLUSIONS

In the past 40 years econometrics has progressed considerably because of the increase in computer power and speed, in data availability, and, possibly consequently, in techniques. The overlap in advanced econometrics textbooks written today is quite small compared with those appearing in the early 1960s, except in the first few chapters. The new techniques have allowed economists to view their data in new ways and sometimes to rephrase theories to link with the techniques and to state the properties of the data differently. What is unclear to me is whether or not the increases in data availability has produced solutions to many economic problems? The possibility of testing international trade theory on data between many countries, which is now possible, should certainly be an advantage, but do we learn a lot from having detailed data from 30 stock markets rather than just the big five, say?

In general, having more data is good. For some problems (e.g., dealing with rare events) economists will continue to need more, given that economies evolve through time, but for many problems there will be reducing returns from increased data, particularly when there is little information available about the quality of the new data relative to the old.

REFERENCES

Finite-Sample Simulation-Based Tests in
Seemingly Unrelated Regressions

Jean-Marie Dufour
Université de Montréal, Montréal, Québec, Canada

Lynda Khalaf
Université Laval, Québec, Québec, Canada

1 INTRODUCTION

In this chapter, we study the problem of testing general, possibly nonlinear constraints on the coefficient of the seemingly unrelated regressions (SURE) model introduced by Zellner (1962). The SURE model may be cast as a system of regression equations with contemporaneously correlated disturbances, where the regressors may differ across equations. For a detailed review, the reader may consult Srivastava and Giles (1987).

In connection with the SURE model, very few analytical finite-sample results are available. A rare exception is provided by Harvey and Phillips (1982, Sect. 3) who derived independence tests between the disturbances of an equation and those of the other equations of a SURE model. The tests involve conventional $F$-statistics and are based on the residuals obtained from regressing each dependent variable on all the independent variables of the system. Of course, this problem is a very special one. In a different vein, Phillips (1985) derived the exact distribution of a two-stage SURE estimator using a fractional matrix calculus. However, the analytical expressions obtained are very complex and, more
importantly, involve unknown nuisance parameters, namely, the elements of the error covariance matrix. The latter fact makes the application of Phillips’ distributional results to practical hypothesis testing problematic.

Asymptotic Wald, Lagrange multiplier, and likelihood ratio tests are available and commonly employed in empirical applications of the SURE model; see, e.g., Breusch (1979) or Srivastava and Giles (1987). It has been shown, however, that, in finite samples, the asymptotic criteria are seriously biased towards over-rejection, with the problem getting worse as the number of equations grows relative to the sample size; see, for e.g., Laitinen (1978), Meisner (1979), Bera, et al. (1981), Theil and Fiebig (1985), and Dufour and Khalaf (2002). Attempts to improve standard asymptotic tests include, in particular: (1) Bartlett-type corrections, and (2) bootstrap and simulation-based methods. See, e.g., Theil and Fiebig (1985), Theil et al. (1985), Taylor et al. (1986), Theil et al. (1986), Rocke (1989), Rayner (1990), and Rilstone and Veall (1996).

Further results relevant to the SURE model can be found in the statistics and econometrics literature on multivariate linear regressions (MLR). These are relevant because the MLR model can be viewed as a special case of the SURE model where the regressor matrices for the different equations are identical. For reviews and further references on exact and asymptotic inference in MLR models, the reader may consult Rao (1973, Chap. 8), Anderson (1984, Chaps. 8 and 13), Kariya (1985), Stewart (1997), and Dufour and Khalaf (2002). In particular, besides showing the inadequacy of various size-correction procedures (including Bartlett corrections) through simulation, we derived in Dufour and Khalaf (2002) exact bounds on the null distribution of LR test statistics for possibly nonlinear hypotheses on regression coefficients in MLR models. Even though computing these bounds analytically may be difficult, they can be easily evaluated by simulation and implemented as finite-sample bounds Monte Carlo (BMC) tests. The implications for hypothesis testing are two-fold. First, the finite-sample bounds on the LR criterion easily yield conservative tests, for both linear and nonlinear hypotheses. Second, Monte Carlo (MC) test methods can lead to tests with correct levels. This is related to the fact that LR statistics are pivotal or boundedly pivotal for quite general hypotheses in the MLR model [see the discussion in Dufour (1997) on boundedly pivotal statistics].

In this chapter, we extend the results presented in Dufour and Khalaf (2002) to the case of SURE systems, under both Gaussian and non-Gaussian disturbance distributions. Indeed the model considered here is an extension of the standard Gaussian SURE model that allows for both Gaussian and non-Gaussian disturbance distribution, as long as the latter is specified up an unknown linear transformation (or contemporaneous
covariance matrix). In particular, we discuss two approaches that can be applied on their own or sequentially, namely: (1) a bounds procedure, and (2) MC tests. Practical implementation of both techniques is simple. To obtain the bounds, we exploit the fact that the SURE specification can be viewed as a special case of a properly chosen MLR model constrained by regressor exclusion restrictions on the different equations.

To be more specific, we give at this point a preliminary discussion of the proposed conservative bound, which can be viewed as an extension of an approach described earlier in Dufour (2002) and Dufour and Kiviet (1998). First, we reconsider the testing problem within the framework of an appropriate MLR model, namely the MLR setup of which the model on hand is a restricted form. As pointed out above, this setup allows for Gaussian and non-Gaussian error distributions, provided the latter can be simulated. Second, we introduce, in the relevant MLR framework, a “uniform linear (UL) hypothesis” [Berndt and Savin (1977)], which is a special case of the set of restrictions specified by the null hypothesis. The intuition behind this suggestion follows from the fact that exact nuisance-parameter free critical values for the LR criterion for testing the suggested UL hypothesis conveniently bounds the LR statistic for testing the general constraints.

In addition, we propose alternative MC tests [see Dwass (1957), Barnard (1963), Jöckel (1986), or Dufour (2002)] that can be run whenever the bounds tests are not conclusive. We consider: (1) an asymptotically valid procedure that may be interpreted as a parametric bootstrap, and (2) a method that is exact for any sample size, following Dufour (2002). Further, in situations where maximum likelihood (ML) methods may be computationally expensive, we introduce LR-type test criteria based on non-ML estimators. In particular, we consider two-stage statistics or estimators at any step of the process by which the likelihood is maximized iteratively. We emphasize that parametric bootstrap and bounds tests should be viewed as complementary rather than as alternative procedures.

The chapter is organized as follows. In Section 2, we present the model studied and define the test statistics that will be considered. In Section 3, we describe the proposed bounds and MC test procedures. Simulation results are reported in Section 4. Section 5 illustrates the procedures proposed by applying them to test restrictions on a factor demand model. We conclude in Section 6.

2 FRAMEWORK

We consider here a $p$-equation SURE system of the form:

$$Y_j = X_j\beta_j + u_j, \quad j = 1, \ldots, p$$

(1)
where $Y_j$ is a vector of $n$ observations on a dependent variable, $X_j$ is a full-column rank $n \times k_j$ matrix of regressors, $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{k_j-1,j})'$ is a vector of $k_j$ unknown coefficients, and $u_j = (u_{1j}, u_{2j}, \ldots, u_{nj})'$ is a $n \times 1$ vector of random disturbances. The system [Eq. (1)] can be rewritten in the stacked form:

$$y = X\beta + u$$

where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}, \quad X = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_p \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p} \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}$$

so $X$ is a $(np) \times k$ matrix, $y$ and $u$ each have dimension $(np) \times 1$, and $\beta$ has dimension $k \times 1$, with $k = \sum_{j=1}^{p} k_j$. Let us also set:

$$U = [u_1, u_2, \ldots, u_p] = \begin{bmatrix} U_{1}^T \\ U_{2}^T \\ \vdots \\ U_{n}^T \end{bmatrix}$$

where $U_{t0} = (u_{t1}, u_{t2}, \ldots, u_{tp})'$ is the disturbance vector for the $t$-th observation.

In the sequel, we shall also use, when required, some or all of the following assumptions and notations:

$$U_{it} = JW_t, \quad t = 1, \ldots, n$$

where $J$ is a fixed lower triangular $p \times p$ matrix such that

$$\Sigma \equiv JJ' = [\sigma_{ij}]_{i,j=1,\ldots,p} \text{ is nonsingular}$$

$W_1, \ldots, W_n$ are $p \times 1$ random vectors

whose joint distribution is completely specified

and

$u$ is independent of $X$.

Assumption (8) is a strict exogeneity assumption, which clearly holds when $X$ is fixed. The assumptions (5)–(7) mean that the disturbance distribution is completely specified up to an unknown linear transformation that can modify the scaling and dependence properties of the disturbances in the different equations. Note that Eqs. (5)–(7) do not necessarily entail that
\( \Sigma \) is the covariance matrix of \( U_t \), because the distribution of \( W_1, \ldots, W_n \) is not restricted (e.g., it may not have finite second moments). However, if we make the additional assumption that

\[
W_1, \ldots, W_n \text{ are uncorrelated, with } \\
E(W_t) = 0, \quad E(W_tW'_t) = I_p, \quad t = 1, \ldots, n
\]

or the stronger assumption:

\[
W_1, \ldots, W_n \overset{i.i.d.}{\sim} N[0, I_p]
\]

we have \( E(U_t, U'_t) = \Sigma, \ t = 1, \ldots, n, \) and

\[
E(u'u') = \Sigma \otimes I_p
\]

Assumption (10) yields the Gaussian SURE model. For further reference, we shall write

\[
W = [W_1, \ldots, W_n]' = UJ
\]

In this chapter, we consider the problem of testing general hypotheses of the form:

\[
H_0: A\beta \in \Delta_0
\]

where \( A \) is a full row-rank \( v_0 \times k \) matrix and \( \Delta_0 \) is a nonempty subset of \( \mathbb{R}^{v_0} \).

For our subsequent arguments, it will be important to spell out the relation between SURE and MLR models. The MLR model may be defined as a SURE model where the regressors in all the equations are the same \((X_1 = X_2 = \cdots = X_p)\). Conversely, a SURE model can be viewed as a restricted MLR system. To be more specific, for each \( 1 \leq j \leq p \) in the context of Eq. (1), let \( \bar{X}_j \) be any matrix such that the columns of \([X_1, \bar{X}_j]\) are linearly independent and span the same space as the columns of the matrix \([X_1, X_2, \ldots, X_p]\). In most practical situations, \( \bar{X}_j \) will simply contain the regressors from the matrices \( X_k, k \neq j \), which are excluded from the \( j \)-th equation. Further, let \( X_s \) be any full-column rank \( n \times k_s \) matrix whose columns span the same space as those of \([X_1, X_2, \ldots, X_p]\), i.e.,

\[
\text{sp}(X_s) = \text{sp}([X_1, X_2, \ldots, X_p]), \quad \det(X'_sX_s) \neq 0, \quad \text{rank}(X_s) = k_s
\]

where, for any matrix \( Z \), \( \text{sp}(Z) \) represents the vector space spanned by the columns of \( Z \). Then, for each \( j \), we can find matrices \( S_j \) and \( \bar{S}_j \) of dimensions \( k_s \times k_j \) and \( k_s \times (k_s - k_j) \), respectively, such that \( X_sS_j = X_j, \ X_s\bar{S}_j = \bar{X}_j \) and the matrix \( T_j = [S_j, \bar{S}_j] \) is invertible. Consequently, Eq. (1) may be rewritten as

\[
Y_j = X_sS_j\beta_j + X_s\bar{S}_j\bar{\beta}_j + u_j = X_s\beta_{sj} + u_j, \quad j = 1, \ldots, p
\]
where $\beta_{s_j} = S_j \beta_j + \tilde{S}_j \tilde{\beta}_j = T_j (\beta_j', \tilde{\beta}_j')'$, with the restrictions:

$$\tilde{\beta}_j = 0, j = 1, \ldots, p \quad (16)$$

The latter restrictions may also be expressed in implicit form on $\beta_{s_j}$, $j = 1, \ldots, p$ as

$$M(S_j) \beta_{s_j} = 0, \quad j = 1, \ldots, p \quad (17)$$

where $M(S_j) = [I_k - S_j (S_j' S_j)^{-1} S_j']$. Clearly Eqs. (15) and (16) define a constrained MLR model. On relaxing the SURE restrictions, Eq. (17), this MLR model can be put in the stacked form:

$$y = (I_p \otimes X_s) \beta_s + u = \tilde{X}_s \beta_s + u \quad (18)$$

where $\tilde{X}_s = I_p \otimes X_s$ and $\beta_s = (\beta_{s_1}, \beta_{s_2}, \ldots, \beta_{s_p})'$, or equivalently,

$$Y = X_s B_s + U \quad (19)$$

where $Y = [Y_1, Y_2, \ldots, Y_p]$ and $B_s = [\beta_{s_1}, \beta_{s_2}, \ldots, \beta_{s_p}]$. We shall call the unrestricted MLR model Eq. (18) [or (19)], an embedding MLR model for the SURE model, Eq. (1). It is clear any hypothesis on $\beta$ can be expressed equivalently in terms of $\beta_s$ within the corresponding embedding MLR model.

In this chapter, we shall emphasize LR-type tests of $H_0$ derived under the Gaussian distributional assumptions, Eqs. (5)–(10). In this case, the log-likelihood function associated with the SURE model [Eq. (2)] has the form:

$$\mathcal{L}(\beta, \Sigma) = -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Sigma|) - \frac{1}{2} (y - X\beta)' (I_n \otimes \Sigma)^{-1} (y - X\beta)$$

(20)

Then, provided that the relevant optima do exist and are unique, the “unconstrained” maximized value of $\mathcal{L}(\beta, \Sigma)$ can be written as

$$L(H_\Sigma) = \sup \{\mathcal{L}(\beta, \Sigma): \beta \in \mathbb{R}^k \text{ and } \Sigma \text{ is p.d.} \}$$

$$= -\frac{np}{2} [\ln(2\pi) + 1] - \frac{n}{2} \ln(|\hat{\Sigma}|) \quad (21)$$

while its constrained maximized value subject to $H_0$ is

$$L(H_0) = \sup \{\mathcal{L}(\beta, \Sigma): A\beta \in \Delta_0 \text{ and } \Sigma \text{ is p.d.} \}$$

$$= -\frac{np}{2} [\ln(2\pi) + 1] - \frac{n}{2} \ln(|\hat{\Sigma}_0|) \quad (22)$$

where $\hat{\Sigma}_0$ and $\hat{\Sigma}_q$ are the restricted and unrestricted ML estimates of $\Sigma$, assuming that $\Sigma$ is positive definite (p.d.). Thus, the Gaussian LR statistic for testing $H_0$ against the unrestricted SURE model [or, equivalently,
against the embedding MLR model, Eq. (18), with the restrictions, Eq. (17) is given by:
\[
LR(H_0) = 2[L(H_S) - L(H_0)] = n \ln(\Lambda_S), \quad \Lambda_S = \frac{\hat{\Sigma}_0}{\hat{\Sigma}_S}
\] (23)

Note that the exclusion SURE restrictions are imposed under both the null and the alternative hypotheses. In the statistics literature, $\Lambda_S^{-1}$ is known as the Wilks criterion.

Similarly, the log-likelihood function associated with the MLR model, Eq. (18), taken jointly with the Gaussian distributional assumptions, Eqs. (5)–(10), is
\[
L(\beta, \Sigma) = - \frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Sigma|) - \frac{1}{2} (y - \hat{X}_s \beta_s)'(\ln \Sigma)^{-1}(y - \hat{X}_s \beta_s)
\] (24)

where $\beta_s = \text{vec}(B_s)$. It is clear that
\[
L(\beta_s, \Sigma) = L(\beta, \Sigma), \text{ when } \hat{\beta}_j = 0, \quad j = 1, \ldots, p
\] (25)

Let
\[
\hat{B}_s = (X'_s X_s)^{-1} X'_s Y, \quad \hat{U} = Y - X_s \hat{B}_s = M(X_s)U, \quad \hat{\Sigma}_M = \frac{1}{n} \hat{U}'\hat{U}
\] (26)

where $M(X_s) \equiv I_n - X_s(X'_s X_s)^{-1} X'_s$. Then, provided that $\hat{U}$ has full-column rank (which requires $n \geq k_s + p), L(\beta_s, \Sigma)$ attains a unique (unconstrained) finite maximum at $\beta_s = \hat{B}_s$ and $\Sigma = \hat{\Sigma}_M$ [see Anderson (1984, Chap. 3)], yielding the maximal value:
\[
L(H_M) = \sup\{L(\beta_s, \Sigma): \beta_s \in \mathbb{R}^{k_s} \text{ and } \Sigma \text{ is p.d.}\}
\] (27)

Thus, $\hat{B}_s$ and $\hat{\Sigma}_M$ are the ML estimators of the parameters of the unrestricted embedding MLR model associated with the SURE model, Eq. (2). Given the assumptions (5)–(10), a necessary and sufficient condition for $\hat{U}$ to have full-column rank with probability one is
\[
P[\text{rank}[M(X_s)W] = p] = 1
\]

The latter will hold, e.g., if $\text{rank}(X_s) = k_s \leq n - p$ and vec($W$) follows an absolutely continuous distribution on $\mathbb{R}^{np}$. In view of the relation between
the SURE model and an embedding MLR model, we shall also consider
the LR statistic for testing \( H_0 \) against the completely unrestricted MLR
model \[\text{Eq. (18) without the SURE restrictions, Eq. (17)}\]:

\[
LR_M(H_0) = 2[L(H_M) - L(H_0)] = n \ln(\Lambda_M), \quad \Lambda_M = |\hat{\Sigma}_0|/|\hat{\Sigma}_M|
\]  

(28)

3 TEST PROCEDURES

We will now show how one can obtain finite-sample LR-based tests in the
context of SURE models as defined above. For that purpose, we shall
exploit special features of so-called uniform linear restrictions for which
LR test statistics have nuisance-parameter null distributions in the context
of MLR models (which entails that they are pivotal statistics under the
null hypothesis). In the MLR case (where \( X_1 = \cdots = X_p \equiv X_s \) and
\( k_1 = \cdots = k_p \equiv k_s \)), the hypothesis \( H_0 \) in Eq. (13) is uniform linear if it
can be expressed in the form:

\[ H_{UL}: (C' \otimes R)\beta = \text{vec}(D_0) \]  

(29)

or, equivalently,

\[ H_{UL}: RBC = D_0 \]  

(30)

where \( B \) is the \( k_s \times p \) matrix such that \( \beta = \text{vec}(B) \), \( R \) is a known \( r \times k_s \) matrix of rank \( r \), \( C \) is a known \( p \times c \) matrix of rank \( c \), and \( D_0 \) is a known \( r \times c \) matrix. In Dufour and Khalaf (2002), it is shown that the null
distribution of the Gaussian LR statistic [derived under the assumptions
(5)–(10)] for testing \( H_{UL} \) (against the unrestricted MLR model) does not
involve any nuisance parameter under the weaker assumptions (1)–(8) —
which allow for non-normal disturbances — and may easily be simulated.
In particular, the parameters of the covariance matrix \( \Sigma = JJ' \) do not
appear in the distribution. Beyond this specific hypothesis class, it is well
known that the LR statistic is not pivotal, even if the null hypothesis is
linear. For further discussion of uniform linear hypotheses in MLR
models, the reader may consult Berndt and Savin (1977), Stewart (1997),
and Dufour and Khalaf (2002).

Let us now turn to the SURE model. In the context of the embedding MLR model, Eq. (18), the null hypothesis \( H_0 \) in Eq. (13) is
equivalent to the conjunction of \( A\beta \in \Delta_0 \) with the SURE restrictions, Eq. (16):

\[ H_0^S: A\beta_s \in \Delta_0 \text{ and } M(S_j)\beta_{sj} = 0, \quad j = 1, \ldots, p \]  

(31)
where \( \beta_s, S_p, M(S), \) and \( k_s \) are defined as in Eqs. (14)–(18), and

\[
F = \begin{bmatrix}
(S'_1 S_1)^{-1} S'_1 & 0 & \cdots & 0 \\
0 & (S'_2 S_2)^{-1} S'_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (S'_p S_p)^{-1} S'_p 
\end{bmatrix}
\]  

(32)

It is clear we can find a full row-rank matrix \( A^* \) of dimension \( v_0 \times (p k_s) \) such that \( H^*_{0 \Lambda} \) can be re-expressed in terms of \( \beta_s \) according to a form similar to \( H^*_0 \) in Eq. (13):

\[
H^*_{0 \Lambda} : A_s \beta_s \in \Delta_{0_\Lambda}
\]

where \( v_0 \geq v_0 \) and \( \Delta_{0_\Lambda} \) is a nonempty subset of \( \mathbb{R}^{p k_s} \).

We now state our main result on the distribution of LR statistics in SURE models.

**Theorem 1** **BOUND ON LR STATISTICS IN SURE MODELS.** Suppose the assumptions (1)–(8) hold, with

\[
P[\text{rank}[M(X_s)W] = p] = 1
\]

(34)

where \( X_s \) is defined as in Eq. (14), and let \( H^*_{UL} : RB_s \Sigma = D_0 \) be a set of uniform linear restrictions on Eq. (19) such that \( H^*_{UL} \) entails \( H^*_{0 \Lambda} \), where the matrices \( R, B_s, \Sigma, D_0, \) and the hypothesis \( H^*_{0 \Lambda} \) are defined as in Eqs. (30) and (31). The following inequalities then hold:

\[
\Lambda^* \leq \Lambda_M \leq \Lambda_{UL}
\]

(35)

where \( \Lambda^* \) and \( \Lambda_M \) are defined as in Eqs. (23) and (28), \( \Lambda_{UL} = |\Sigma_{UL}| / |\Sigma_M| \) and \( \Sigma_{UL} \) is a ML estimator of \( \Sigma \) obtained under the uniform linear restrictions \( H^*_{UL} \). Furthermore, under \( H^*_{UL} \), the distribution of \( \Lambda_{UL} \) (conditional on \( X \)) does not depend on the unknown parameter matrices \( B \) and \( \Sigma \) nor on the values of the constants in \( D_0 \).

**Proof.** The proof is based on observing that the hypotheses involved in the definitions of the statistics \( \Lambda^*, \Lambda_M, \) and \( \Lambda_{UL} \) can be viewed as special cases of the embedding MLR model, Eq. (19). First, we note that, under the assumptions (1)–(8) and (34), the log-likelihood \( \mathcal{L}_s(\beta_s, \Sigma) \) has (with probability one) a unique maximum given by \( L(H_M) \) in Eq. (27). This entails that the supremum of \( \mathcal{L}_s(\beta_s, \Sigma) \) under any set of restrictions on \( \beta_s \) must be finite. In particular, the supremum (with respect to \( \beta_s \) and \( \Sigma \)) of \( \mathcal{L}_s(\beta_s, \Sigma) \) under \( H^*_{UL} \) can be written

\[
L(H^*_{UL}) = \sup \{\mathcal{L}_s(\beta_s, \Sigma) : RB_s \Sigma = D_0 \text{ and } \Sigma \text{ is p.d.}\}
\]

\[
= -\frac{n p}{2}[\ln(2\pi) + 1] - \frac{n}{2}\ln(|\Sigma_{UL}|)
\]

(36)

\[ \]
where $\hat{\Sigma}_{ul}$ is the ML estimator of $\Sigma$ under $H_{ul}^0$. Thus, the Gaussian LR statistic for testing $H_{ul}^0$ against the embedding MLR model [i.e., Eq. (18) jointly with Eqs. (5)-(10)] is

$$LR_M(H_{ul}^0) = 2[L(H_M) - L(H_{ul}^0)] = n \ln(\Lambda_{ul}),$$

$$\Lambda_{ul} = |\hat{\Sigma}_{ul}|/|\Sigma|$$

From Theorem 3.1 in Dufour and Khalaf (2002), the exact distribution of $\Lambda_{ul}$ under $H_{ul}^0$ only depends on the distribution of $W$ and the known matrices $X$, $R$, and $C$, but not on $D_0$ nor on the otherwise unknown parameters in $B$ and $\Sigma$.

Now, by the definition of the embedding MLR model, Eq. (25), and the equivalence between $H_0$ and $H_{ul}^0$, we see that $L(H_S)$ and $L(H_0)$ in Eqs. (21)-(22) can also be expressed in terms of $L(\hat{\Sigma}_{ul})$:

$$L(H_S) = \sup \{G(\beta, \Sigma): M(S)\beta_{sj} = 0, \quad j = 1, \ldots, p, \text{ and } \Sigma \text{ is p.d.}\}$$

$$L(H_0) = \sup \{G(\beta, \Sigma): \beta_s \text{ satisfies } H_0^0 \text{ and } \Sigma \text{ is p.d.}\}$$

Since $H_{ul}^0$ entails $H_0^0$, which in turn is a restricted form of the SURE model $(HS)$, and since $H_S$ can be obtained by imposing linear restrictions on the embedding MLR model $(H_M)$, it follows that

$$L(H_{ul}^0) \leq L(H_0) \leq L(H_S) \leq L(H_M)$$

hence

$$L(H_S) - L(H_0) \leq L(H_M) - L(H_0) \leq L(H_M) - L(H_{ul}^0)$$

and

$$\Lambda_S \leq \Lambda_M \leq \Lambda_{ul}$$

This completes the proof of the theorem.

It follows from the latter theorem that

$$P[\Lambda_S \geq \lambda_{ul}(\alpha)] \leq P[\Lambda_M \geq \lambda_{ul}(\alpha)] \leq \alpha$$

under $H_0$, where $\lambda_{ul}(\alpha)$ is determined such that $P[\Lambda_{ul} \geq \lambda_{ul}(\alpha)] = \alpha$ (or, at least, $P[\Lambda_{ul} \geq \lambda_{ul}(\alpha)] \leq \alpha$) and $0 \leq \alpha \leq 1$. It is important to note here that the inequality (35) holds for any set of uniform linear restrictions which entails $H_0$. In particular, on taking $R = I_k$, $C = I$, and $D_0 = B_0$ (the true value of $B_0$), it is clear such a set of restrictions does always exist, although other choices (for $R$, $C$, and $D_0$) may be available in view of the form of $H_0$ and lead to a tighter bound. Further, the distribution of $\Lambda_{ul}$ (when $RB_0C = D_0$) only depends on $X$, $R$, and $C$, but not on $D_0$, so one
can use any possible value of \(D_0\) (such as \(D_0 = 0\)) in order to compute or simulate this distribution.

Using well known results from Anderson (1984) and Rao (1973) for Gaussian MLR models, it is also possible to show that the bounding statistic \(\Lambda_{\text{UL}}\) is distributed like the product of the inverse of \(p\) beta variables with degrees of freedom that depend only on \(r, c, p,\) and \(k^*\) [see Dufour (1997) and Dufour and Khalaf (2002)]. The latter result is, however, hardly useful for practical applications of the proposed bound. Hence, we do not restate our conclusions here for this specific Gaussian case: it is more convenient to derive \(\Lambda_{\text{UL}}(\alpha)\) by simulation as shown below and using Theorem 1, under any distributional assumptions that satisfy Eq. (5) including the normal case. Finally, we note that the same bound applies to both criteria \(\Lambda_s\) and \(\Lambda_M\). Since \(\Lambda_M \geq \Lambda_S\), it will thus be preferable to apply the bound to \(\Lambda_M\) rather than \(\Lambda_S\), since this will yield a more powerful test.

Theorem 1 has further implications for LR-based hypothesis tests. The fact that the null distribution of the LR statistic can be bounded (in a nontrivial way) by a statistic whose distribution can be simulated fairly easily entails that MC test techniques may be used to obtain finite-sample \(p\)-values based on the LR-based statistics when the bounds test is not conclusive. In earlier work, we have discussed in detail how such procedures can be implemented; see Dufour and Kiviet (1996, 1998), Dufour (2002), Dufour et al. (1998), and Dufour and Khalaf (2001, 2002). These include techniques for the construction of: (1) a (parametric) bootstrap-type \(p\)-value which we denote a local MC \(p\)-value (LMC) to account for the fact that the underlying simulation routine is implemented given a specific nuisance parameter estimate, and (2) an exact randomized \(p\)-value that corresponds to the largest MC \(p\)-value over the relevant nuisance parameter space; conformably, we call the latter a maximized Monte Carlo (MMC) \(p\)-value [Dufour (2002)]. Both procedures are summarized below. Although the LMC \(p\)-value is only valid asymptotically, nonrejections are conclusive from a finite-sample perspective, in the following sense. Indeed, for all \(0 \leq \alpha \leq 1\), if the LMC \(p\)-value exceeds \(\alpha\), we can be sure that the maximum \(p\)-value also exceeds \(\alpha\). We emphasize the fact that the MMC test can be implemented in complementarity with the above defined bounds tests. Indeed, if the bounds test rejects the null then the MMC test is certainly significant. For a more detailed discussion of the justification and implementation of such simulation-based procedures, we refer the reader to the papers just cited.

To illustrate how the above results may be used in the context of a SURE model, we will now discuss an illustrative example.
Example 1 Three-equation SURE model. In the SURE model, Eq. (1), with Gaussian errors and \( p = 3, k_i = 2, X_i = [t_{ir}, x_i] \) where \( t_r \) denotes a vector of \( n \) 1’s, consider the problem of testing

\[
H_0: \beta_{11} = \beta_{22} = \beta_{33}
\]  
(44)

We suppose also that the matrix \( X_s = [t_r, x_1, x_2, x_3] \) has full-column rank \( k_s = 4 \leq n - 3 \). Then, it is easy to see that this problem is equivalent to testing

\[
H_0^\prime: b_{11} = b_{22} = b_{33} \quad \text{and} \quad b_{12} = b_{13} = b_{21} = b_{23} = b_{31} = b_{32} = 0
\]

in the framework of the MLR model:

\[
Y = X_s B_s + U
\]  
(45)

with \( Y = [Y_1, Y_2, Y_3], U = [U_1, U_2, U_3], B_s = [b_1, b_2, b_3], \) and \( b_j = (b_{0j}, b_{1j}, b_{2j}, b_{3j}), j = 1, 2, 3 \). In order to use the above results on the conservative bound, we need to construct a set of UL restrictions on the coefficients of the later MLR model that satisfy the hypothesis in question. It is easy to see that constraints setting the values of the coefficients \( b_{ij}, i, j = 1, \ldots, 3 \) to specific values (9 restrictions on 12 coefficients) meet this purpose:

\[
H_{UL}^\prime:
\begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
B_s = D_0
\]  
(46)

All that remains is to calculate the LR, as defined in Eq. (23), and use the critical value associated with the uniform linear restriction, Eq. (46).

A bounds MC test may then be applied as described in Dufour and Khalaf (2002) for testing general possibly nonlinear restrictions in MLR models. The procedure can be described as follows for the example just considered. For further reference, and conformably with Dufour and Khalaf (2002), we call the latter bounds test, a BMC test.

1. Denote \( \Lambda^{(0)} \) the observed test statistic, which could be the statistic \( \Lambda_s \) in Eq. (23) or \( \Lambda_M \) in Eq. (28).

2. By MC methods, draw \( N \) simulated samples (conditional on the right-hand-side regressors) from model (18) in \( H_{UL}^\prime \). For instance, in the above example, one way to do this is to draw the simulated samples from the base model, Eq. (45), with parameters set to their constrained (imposing \( H_0^\prime \)) SURE estimates on which the additional restrictions underlying \( H_{UL}^\prime \) have been imposed (simply setting to fixed values the relevant coefficients of \( B_s \)). To be more specific, let us denote by \( \tilde{b}_{ij} \) the SURE estimate of \( \beta_{ij} \), \( 0 \leq i \leq 3, 1 \leq j \leq 3 \), several of which should be zero to account for the SURE exclusion restrictions. Then, if we choose to draw
from model (45) with coefficients $\tilde{b}_{ij}, i = 0, \ldots, 3, j = 1, \ldots, 3$ and the conformable covariance matrix estimate, then $H_{UL}$ could be of the form: $b_{ij} = \tilde{b}_{ij}, i, j = 1, \ldots, 3$.* Furthermore, any error distribution that satisfies Eqs. (5)–(8) may be considered.

3. From each simulated sample, compute the bounding statistic $\Lambda_{UL}$, which corresponds to the LR-based statistic associated with $H_{UL}$: this yields $\Lambda_{UL}^{(h)}, h = 1, \ldots, N$. As emphasized above, it is important to make sure that the regression coefficients selected to generate the MC drawings correspond to the restrictions implied by $H_{UL}$. In other words, the simulated values of the bounding statistic should satisfy the null hypothesis.

4. Compute the simulated $p$-value $\tilde{p}_N(\Lambda^{(0)})$ where

$$\tilde{p}_N(x) = \left\{ 1 + \sum_{h=1}^{N} I[\Lambda_{UL}^{(h)} - x] \right\} / (N + 1)$$

(47)

$I[z] = 1$ if $z \geq 0$ and $I[z] = 0$ if $z < 0$.

5. The procedure rejects at level $\alpha$ if $\tilde{p}_N(\Lambda^{(0)}) \leq \alpha$.

Using the same arguments as in Dufour and Kiviet (1996, 1998), Dufour (2002), and Dufour and Khalaf (2002), it is easy to see that

$$P[\tilde{p}_N(\Lambda^{(0)}) \leq \alpha] \leq \alpha \text{ under } H_0$$

so the critical region $\tilde{p}_N(\Lambda^{(0)}) \leq \alpha$ has level $\alpha$. If the above procedure is implemented replacing $\Lambda_{UL}^{(h)}, j = 1, \ldots, N$ with $\hat{\Lambda}^{(h)}, h = 1, \ldots, N$, which refer to realized values of the LR criterion associated with $H_0$ and the simulated samples, then Eq. (47) yields a parametric bootstrap or an LMC $p$-value. In this case, the $p$-value in question depends on the choice of the intervening nuisance parameters. The MMC $p$-value corresponds to the largest MC $p$-value overall nuisance parameters compatible with the null hypothesis. A global optimizing algorithm is required to maximize the MC $p$-value. In this chapter, we used simulated annealing [see Goffe et al. (1994)]. The reader may consult Dufour and Kiviet (1996, 1998), Dufour (2002), Dufour et al. (1998), and Dufour and Khalaf (2001) for further discussion of MC tests in econometrics.

*It is also possible, and perhaps more efficient, to rewrite the bounding LR statistic as a pivotal quantity, using the results in Dufour and Khalaf (2002). The procedure just presented exploits the pivotal property of the statistic implicitly. Yet it is quite intuitive and relates to the familiar parametric bootstrap.
4 SIMULATION STUDY

In this section, we present simulation results illustrating the performance of the above proposed procedures along with the one of more traditional asymptotically justified methods. In particular, we examine the performance of BMC and MMC tests in SURE contexts. In the linear case, we also consider LMC tests based on standard Wald-type criteria and several alternative statistics justified on the basis of computational cost as opposed to those relying on full maximum likelihood estimation.

4.1 Design

We studied two Gaussian SURE designs, similar to that of Example 1. In the first one (D1), we considered the problem of testing linear cross-equation constraints, while in the second one (D2), we studied a nonlinear constraint.

D1. SURE system, cross-equation constraints

Model (1); \( k_j = 2, \ j = 1, \ldots, p; \ p = 3, 5; \ n = 25; \)

\( H_0: \beta_{ij} = \beta_{1j}, \ j = 2, \ldots, p. \)

D2. SURE system, non-linear constraints

Model (1); \( k_j = 3, \ j = 1, \ldots, p; \ p = 7; \ n = 25; \)

the regressor which corresponds to \( \beta_{ij} \) is common to all equations;

\( H_0: \beta_{ij} = \gamma \beta_{2j}, \ j = 1, \ldots, p. \ \gamma \) unknown.

For each model, a constant regressor was included and the other regressors were independently drawn (once) from a normal distribution; the errors were independently generated as i.i.d. \( N(0, \Sigma) \) with \( \Sigma = JJ' \) and the elements of \( J \) drawn (once) from a normal distribution. The coefficients are reported in Table 1.

The statistics examined for D1 include the relevant LR criteria defined by Eqs. (23) and (28), as well as three other types of statistics: (1) quasi-LR statistics based on incompletely maximized likelihood functions;

| TABLE 1 Coefficient Values used in the Simulation Experiments |
|---------------|-----------------|
| D1            | \( \beta_{\text{3ED}} = (1.2, 0.1, 0.8, 0.1, -1.1, 0.1) \)’ |
|               | \( \beta_{\text{4ED}} = (1.2, 0.1, 0.8, 0.1, -1.1, 0.1, 1.9, 0.1, -0.2, 0.1) \)’ |
| D2            | \( \gamma = 0.009 \) and \( \beta_{g}, \beta_{2j}, j = 1, \ldots, p. \) drawn (once) as i.i.d. \( N(0, 0.16) \) |
(2) test statistics similar to those suggested by Theil et al. (1985), and (3) a number of Wald-type criteria. To be more precise, the latter are defined as follows:

1. The quasi-LR (QLR) statistics are

\[
QLR_{(l)} = n \ln(\Lambda_{(l)}), \quad \Lambda_{(l)} = \frac{|\hat{\Sigma}_{(l)}|}{|\hat{\Sigma}_{0(0)}|}
\]  

(48)

where \( \hat{\Sigma}_{0(0)} \) and \( \hat{\Sigma}_{(l)} \) denote the constrained and unconstrained iterative estimators of \( \Sigma \) and the subscript \( l \) refers to the number of iterations involved. Though we did not analytically establish the asymptotic distribution of the latter criteria, we assessed their asymptotic significance using the \( \chi^2 \) reference distribution for the usual LR statistic. We append the subscript \( LMC \) to the notation for the QLR test to refer to the corresponding LMC test.

2. The test statistics suggested by Theil et al. (1985) may be interpreted as unscaled Wald-type statistics, whose level is controlled by a MC (or bootstrap) method. We consider these here mainly for historical reasons, because they are really the first simulation-based test procedures proposed in the SURE setup. For the model with three equations, we considered:

\[
\mu_{31} = |\hat{\beta}_{11} - \hat{\beta}_{22}| + |\hat{\beta}_{22} - \hat{\beta}_{33}|
\]
\[
\mu_{32} = |\hat{\beta}_{11} - \hat{\beta}_{33}| + |\hat{\beta}_{22} - \hat{\beta}_{33}|
\]
\[
\mu_{33} = |\hat{\beta}_{11} - \hat{\beta}_{22}| + |\hat{\beta}_{11} - \hat{\beta}_{33}|
\]

In the five-equation case, the following were selected among many possible choices:

\[
\mu_{51} = |\hat{\beta}_{11} - \hat{\beta}_{22}| + |\hat{\beta}_{22} - \hat{\beta}_{33}| + |\hat{\beta}_{33} - \hat{\beta}_{44}| + |\hat{\beta}_{44} - \hat{\beta}_{55}|
\]
\[
\mu_{52} = |\hat{\beta}_{22} - \hat{\beta}_{33}| + |\hat{\beta}_{33} - \hat{\beta}_{44}| + |\hat{\beta}_{44} - \hat{\beta}_{55}| + |\hat{\beta}_{55} - \hat{\beta}_{11}|
\]
\[
\mu_{53} = |\hat{\beta}_{33} - \hat{\beta}_{44}| + |\hat{\beta}_{44} - \hat{\beta}_{55}| + |\hat{\beta}_{55} - \hat{\beta}_{11}| + |\hat{\beta}_{11} - \hat{\beta}_{22}|
\]
\[
\mu_{54} = |\hat{\beta}_{44} - \hat{\beta}_{55}| + |\hat{\beta}_{55} - \hat{\beta}_{11}| + |\hat{\beta}_{11} - \hat{\beta}_{22}| + |\hat{\beta}_{22} - \hat{\beta}_{33}|
\]
\[
\mu_{55} = |\hat{\beta}_{55} - \hat{\beta}_{11}| + |\hat{\beta}_{11} - \hat{\beta}_{22}| + |\hat{\beta}_{22} - \hat{\beta}_{33}| + |\hat{\beta}_{33} - \hat{\beta}_{44}|
\]

For the purpose of this experiment, we used for \( \hat{\beta} \) the Gaussian ML estimator of \( \beta \).

3. The Wald-type criteria are based on feasible generalized least squares (GLS) parameter estimates. Specifically, we considered the statistic suggested in Srivastava and Giles (1987, Chap. 10) for
an hypothesis of the form $A\beta - r = 0$:

$$W = \frac{v_1}{v_0} \frac{(A\hat{\beta} - r)'(A(S^{-1} \otimes I_p)X)^{-1}A'(A\hat{\beta} - r)}{(y - X\hat{\beta})'(S^{-1} \otimes I_p)(y - X\hat{\beta})}$$

(49)

where $A$ is a $v_0 \times k$ full-row rank fixed matrix, $v_1 = np - k$, $k = \sum_{i=1}^{p} k_i$, while $S$ and $\hat{\beta}$ are feasible generalized least squares parameter estimates. Under the null hypothesis (and standard regularity conditions), $v_0 W$ has a $\chi^2$ asymptotic distribution. Theil (1971, Chap. 6) suggests that the $F(v_0, v_1)$ distribution better captures the finite-sample distribution of the statistics. Yet this claim is not supported by either analytical or simulation evidence. Maximum likelihood estimators may also be substituted for $\hat{\beta}$ and $S$ in the formulas for the Wald criterion.

Here, we have considered both the standard feasible estimator of $\beta$ [using the estimate of $S$ based on least squares residuals applied to each one of the regressions in Eq. (1)] as well as the ML estimators (iterated to convergence) of $\beta$ and $\Sigma$. For further reference, the GLS and ML based $W$ tests will be denoted $W_{\text{GLS}, j}$ and $W_{\text{ML}, j}$, respectively, where $j \in \{\chi^2, F\}$ indicates whether the critical value was obtained from the $\chi^2$ asymptotic distribution or from the $F(v_0, v_1)$ distribution. The local MC (parametric bootstrap) counterparts will be denoted $W_{\text{GLS}, \text{LMC}}$ and $W_{\text{ML}, \text{LMC}}$.

In both D1 and D2, we computed empirical frequencies of type I errors, based on a nominal size of 5% and 1000 replications. In D1, the powers of the tests were investigated by simulating the model with the same parameter values except for $\beta_{11}$ respectively. The LMC and BMC tests were applied with 19 and 99 replications. Because of the computational cost involved, the MMC test was only applied with $p = 3$ and $N = 19$. The BMC test was performed based on the bounding statistic as described in Example 1. For each test statistic, the LMC randomized procedure was based on simulations that use a restricted estimator similar to the estimator(s) involved in the corresponding test statistic: a restricted ML (or quasi-ML) estimator for LR or Wald-type tests based on ML (quasi-ML) estimators, and restricted feasible GLS estimators for tests based on GLS estimators. All the experiments were conducted using Gauss-386i VM version 3.1.

*The statistic $W$ above corresponds to the $z$ statistic in Eq. (10.11) of Srivastava and Giles (1987, Chap. 10).

1For the purpose of power comparisons, the asymptotic tests were size corrected using an independent simulation.
4.2 Results and Discussion

The results of the limited size study in D2 reveal the following: the observed empirical frequency of type I errors for the LR statistic was 12.5% whereas the bounds test (2.6%) satisfied the 5% level constraint. The results of experiment D1 are summarized in Tables 2–5. The subscripts \( \text{asy} \), \( \text{BMC} \), \( \text{LMC} \), and \( \text{MMC} \), which appear in these tables, refer, respectively, to the standard asymptotic tests, MC bounds tests, local MC tests (parametric bootstrap), and maximized MC tests. \( LR[\text{asy}] \),

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Empirical Levels of Various Tests: Experiment D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic tests</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>3EQ</td>
</tr>
<tr>
<td>( \mathcal{W}[\text{GLS}, \chi^2] )</td>
<td>0.061</td>
</tr>
<tr>
<td>( \mathcal{W}[\text{ML}, \chi^2] )</td>
<td>0.124</td>
</tr>
<tr>
<td>( \mathcal{W}[\text{GLS}, F] )</td>
<td>0.052</td>
</tr>
<tr>
<td>( \mathcal{W}[\text{ML}, F] )</td>
<td>0.111</td>
</tr>
<tr>
<td>( LR[\text{asy}] )</td>
<td>0.094</td>
</tr>
<tr>
<td>( QLR(0) )</td>
<td>0.068</td>
</tr>
<tr>
<td>( QLR(1) )</td>
<td>0.088</td>
</tr>
<tr>
<td>( QLR(2) )</td>
<td>0.094</td>
</tr>
<tr>
<td>( QLR(2)_5 )</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Note: \( p_1 \) is the empirical probability that \( LR[\text{LMC}] \) and \( LR[\text{BMC}] \) reject, \( p_2 \) measures the probability that \( LR[\text{BMC}] \) fails to reject and \( LR[\text{LMC}] \) rejects, \( p_3 \) measures the probability that \( LR[\text{BMC}] \) rejects and \( LR[\text{LMC}] \) fails to reject, and \( p_4 \) is the empirical probability that both tests fail to reject.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Power of the Bounds Tests: Experiment D1 (( H_0 : \beta_{11} = 0.1 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( \beta_{11} )</td>
</tr>
<tr>
<td>19 ( p_1 )</td>
<td>0.065</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0.171</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>0.030</td>
</tr>
<tr>
<td>( p_4 )</td>
<td>0.734</td>
</tr>
<tr>
<td>99 ( p_1 )</td>
<td>0.077</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0.204</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>0.022</td>
</tr>
<tr>
<td>( p_4 )</td>
<td>0.697</td>
</tr>
</tbody>
</table>

Note: \( p_1 \) is the empirical probability that \( LR[\text{LMC}] \) and \( LR[\text{BMC}] \) reject, \( p_2 \) measures the probability that \( LR[\text{BMC}] \) fails to reject and \( LR[\text{LMC}] \) rejects, \( p_3 \) measures the probability that \( LR[\text{BMC}] \) rejects and \( LR[\text{LMC}] \) fails to reject, and \( p_4 \) is the empirical probability that both tests fail to reject.
<table>
<thead>
<tr>
<th>$\beta_{11}$</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.0</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W[GLS, \chi^2]$</td>
<td>0.192</td>
<td>0.647</td>
<td>0.939</td>
<td>0.999</td>
<td>0.999</td>
<td>0.192</td>
<td>0.647</td>
<td>0.939</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>$W[ML, \chi^2]$</td>
<td>0.264</td>
<td>0.787</td>
<td>0.984</td>
<td>1.000</td>
<td>1.000</td>
<td>0.264</td>
<td>0.787</td>
<td>0.984</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(\text{asy})$</td>
<td>0.281</td>
<td>0.806</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
<td>0.281</td>
<td>0.806</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$W[GLS, \text{LMC}]$</td>
<td>0.185</td>
<td>0.579</td>
<td>0.884</td>
<td>0.974</td>
<td>0.986</td>
<td>0.202</td>
<td>0.640</td>
<td>0.934</td>
<td>0.990</td>
<td>0.998</td>
</tr>
<tr>
<td>$W[ML, \text{LMC}]$</td>
<td>0.225</td>
<td>0.704</td>
<td>0.958</td>
<td>0.997</td>
<td>1.000</td>
<td>0.260</td>
<td>0.774</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(\text{asy})$</td>
<td>0.236</td>
<td>0.707</td>
<td>0.962</td>
<td>0.997</td>
<td>1.000</td>
<td>0.262</td>
<td>0.779</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$W[GLS, \text{LMC}]$</td>
<td>0.185</td>
<td>0.579</td>
<td>0.884</td>
<td>0.974</td>
<td>0.986</td>
<td>0.202</td>
<td>0.640</td>
<td>0.934</td>
<td>0.990</td>
<td>0.998</td>
</tr>
<tr>
<td>$W[ML, \text{LMC}]$</td>
<td>0.225</td>
<td>0.704</td>
<td>0.958</td>
<td>0.997</td>
<td>1.000</td>
<td>0.260</td>
<td>0.774</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(\text{asy})$</td>
<td>0.236</td>
<td>0.707</td>
<td>0.962</td>
<td>0.997</td>
<td>1.000</td>
<td>0.262</td>
<td>0.779</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$QLR(0)$</td>
<td>0.227</td>
<td>0.689</td>
<td>0.950</td>
<td>0.993</td>
<td>0.998</td>
<td>0.256</td>
<td>0.762</td>
<td>0.977</td>
<td>0.997</td>
<td>0.998</td>
</tr>
<tr>
<td>$QLR(1)$</td>
<td>0.238</td>
<td>0.709</td>
<td>0.961</td>
<td>0.997</td>
<td>1.000</td>
<td>0.259</td>
<td>0.776</td>
<td>0.986</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$QLR(2)$</td>
<td>0.236</td>
<td>0.707</td>
<td>0.962</td>
<td>0.997</td>
<td>1.000</td>
<td>0.262</td>
<td>0.779</td>
<td>0.985</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(BMC)$</td>
<td>0.197</td>
<td>0.552</td>
<td>0.869</td>
<td>0.974</td>
<td>0.992</td>
<td>0.210</td>
<td>0.641</td>
<td>0.935</td>
<td>0.995</td>
<td>0.998</td>
</tr>
<tr>
<td>$\mu_{31}$</td>
<td>0.076</td>
<td>0.108</td>
<td>0.148</td>
<td>0.216</td>
<td>0.259</td>
<td>0.064</td>
<td>0.108</td>
<td>0.165</td>
<td>0.219</td>
<td>0.268</td>
</tr>
<tr>
<td>$\mu_{32}$</td>
<td>0.197</td>
<td>0.552</td>
<td>0.869</td>
<td>0.974</td>
<td>0.992</td>
<td>0.210</td>
<td>0.641</td>
<td>0.935</td>
<td>0.995</td>
<td>0.998</td>
</tr>
<tr>
<td>$\mu_{33}$</td>
<td>0.093</td>
<td>0.183</td>
<td>0.307</td>
<td>0.432</td>
<td>0.489</td>
<td>0.088</td>
<td>0.184</td>
<td>0.328</td>
<td>0.503</td>
<td>0.601</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta_{11}$</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W[GLS, \chi^2]$</td>
<td>0.200</td>
<td>0.703</td>
<td>0.961</td>
<td>0.994</td>
<td>0.999</td>
<td>0.200</td>
<td>0.703</td>
<td>0.961</td>
<td>0.994</td>
<td>0.999</td>
</tr>
<tr>
<td>$W[ML, \chi^2]$</td>
<td>0.317</td>
<td>0.918</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.317</td>
<td>0.918</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(\text{asy})$</td>
<td>0.331</td>
<td>0.913</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>0.331</td>
<td>0.913</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$W[GLS, \text{LMC}]$</td>
<td>0.162</td>
<td>0.619</td>
<td>0.918</td>
<td>0.982</td>
<td>0.998</td>
<td>0.186</td>
<td>0.684</td>
<td>0.946</td>
<td>0.990</td>
<td>0.999</td>
</tr>
<tr>
<td>$W[ML, \text{LMC}]$</td>
<td>0.265</td>
<td>0.832</td>
<td>0.991</td>
<td>0.999</td>
<td>1.000</td>
<td>0.297</td>
<td>0.903</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(\text{asy})$</td>
<td>0.286</td>
<td>0.841</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.328</td>
<td>0.908</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$QLR(0)$</td>
<td>0.225</td>
<td>0.849</td>
<td>0.988</td>
<td>0.998</td>
<td>1.000</td>
<td>0.290</td>
<td>0.900</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$QLR(1)$</td>
<td>0.256</td>
<td>0.860</td>
<td>0.971</td>
<td>0.998</td>
<td>1.000</td>
<td>0.316</td>
<td>0.864</td>
<td>0.983</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>$QLR(2)$</td>
<td>0.287</td>
<td>0.842</td>
<td>0.991</td>
<td>0.999</td>
<td>1.000</td>
<td>0.331</td>
<td>0.908</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$LR(BMC)$</td>
<td>0.120</td>
<td>0.427</td>
<td>0.792</td>
<td>0.958</td>
<td>0.995</td>
<td>0.110</td>
<td>0.484</td>
<td>0.877</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>$\mu_{31}$</td>
<td>0.029</td>
<td>0.034</td>
<td>0.038</td>
<td>0.041</td>
<td>0.048</td>
<td>0.032</td>
<td>0.036</td>
<td>0.039</td>
<td>0.041</td>
<td>0.044</td>
</tr>
<tr>
<td>$\mu_{32}$</td>
<td>0.031</td>
<td>0.036</td>
<td>0.039</td>
<td>0.042</td>
<td>0.045</td>
<td>0.031</td>
<td>0.034</td>
<td>0.038</td>
<td>0.040</td>
<td>0.041</td>
</tr>
<tr>
<td>$\mu_{33}$</td>
<td>0.042</td>
<td>0.086</td>
<td>0.154</td>
<td>0.258</td>
<td>0.359</td>
<td>0.035</td>
<td>0.077</td>
<td>0.152</td>
<td>0.241</td>
<td>0.397</td>
</tr>
<tr>
<td>$\mu_{44}$</td>
<td>0.023</td>
<td>0.071</td>
<td>0.159</td>
<td>0.289</td>
<td>0.456</td>
<td>0.025</td>
<td>0.067</td>
<td>0.175</td>
<td>0.302</td>
<td>0.512</td>
</tr>
<tr>
<td>$\mu_{55}$</td>
<td>0.031</td>
<td>0.050</td>
<td>0.071</td>
<td>0.118</td>
<td>0.170</td>
<td>0.033</td>
<td>0.056</td>
<td>0.092</td>
<td>0.128</td>
<td>0.180</td>
</tr>
</tbody>
</table>
\(LR[BM\text{C}], LR[LM\text{C}],\) and \(LR[MM\text{C}]\) refer to the corresponding LR tests, and similarly to \(LR_M\) and \(QLR\). Our results show the following:

1. The asymptotic criteria have an upward bias in size; as can be seen in Table 2, rejection of the null is repeatedly many times larger than what it should be. The bias clearly worsens in the five equation example (5EQ). Across the cases examined, the Wald-type statistics have larger sizes when based on their asymptotic \(\chi^2\) critical values. Although the \(F\) approximation seems to correct the problem in the 3EQ model, it clearly fails to do so in the 5EQ case. The nonlinear LR test examined in D2 is also oversized.

2. The BMC test was found to be well behaved. Power gains are possible in other test problems where a tighter critical bound is available. Indeed, we have observed reasonable power even if we have experimented with the worst scenario, in the sense that bounding test statistics correspond to a null hypothesis that fixes the values of all regression coefficients (except the intercept). Furthermore, we found that the BMC and the MMC tests based on \(LR_M\) yield equivalent decisions for all cases examined; the MMC test based on \(LR\) performs marginally better. This illustrates the value of the conservative bounds test as a tool to be used in conjunction with LMC test methods and not necessarily as an alternative to those methods. As emphasized earlier, the bounds procedure is computationally inexpensive and exact. In addition, whenever the bounds test rejects, inference may be made without further appeal to randomized tests.

3. There is no indication of over-rejection for the LMC tests considered. While the critical values used, conditional on the particular choice of consistent estimator for the error covariance matrix, were only asymptotically justified, the procedure was remarkably effective in correcting the bias. Whether this conclusion would carry to larger systems remains an open question. In this regard, note that available simulation evidence on the SURE model, specifically the experiment in Rocke (1989) on large systems, is limited to three equations at best.

4. While they did exhibit adequate sizes, the statistics inspired by Theil et al. (1985) did not fare well in terms of power. For the 3EQ model, the performance was dramatically poor for \(\mu_{32}\) and \(\mu_{33}\) but less so in the case of \(\mu_{31}\). Even then, as compared to the randomized LR, the performance is less than satisfactory.

5. The LMC tests performed noticeably well in terms of power in all instances, even when the number of replications was as low as 19.
It is worth noting, however, that simulation evidence does not favor the randomized usual LR tests over those based on $\Lambda(ij)$ typically involving fewer iterations, although we are uncertain as to the asymptotic equivalence of both procedures. This observation has an important bearing on empirical practice. The simplicity of the method based on $\Lambda(ij)$ has much to recommend it for larger models in which statistics requiring full ML may be quite expensive to randomize.

5 EMPIRICAL ILLUSTRATION

In this section, we present an empirical application that illustrates the results presented in this paper. We consider testing restrictions on the parameters of a generalized Leontief cost function. We used the data from Berndt and Wood (1975) and the factor demand system from Berndt (1991, pp. 460–462). The model imposes constant returns to scale and linear homogeneity in prices, and includes four inputs: capital (K), labor (L), energy (E), and nonenergy intermediate materials (M). If we denote the output by $Y$ and the input prices by $P_j$, $j = K, L, E, M$, the stochastic cost minimizing input–output KLEM equations are

\begin{align*}
K/Y &= d_{KK} + d_{KL}(P_L/P_K)^{1/2} + d_{KE}(P_E/P_K)^{1/2} + d_{KM}(P_M/P_K)^{1/2} + e_K \\
L/Y &= d_{LL} + d_{LK}(P_K/P_L)^{1/2} + d_{LE}(P_E/P_L)^{1/2} + d_{LM}(P_M/P_L)^{1/2} + e_L \\
E/Y &= d_{EE} + d_{EK}(P_K/P_E)^{1/2} + d_{EL}(P_L/P_E)^{1/2} + d_{EM}(P_M/P_E)^{1/2} + e_E \\
M/Y &= d_{MM} + d_{MK}(P_K/P_M)^{1/2} + d_{ML}(P_L/P_M)^{1/2} + d_{ME}(P_E/P_M)^{1/2} + e_M
\end{align*}

where the error terms $e_K, e_L, e_E, e_M$ satisfy the distributional assumptions, Eq. (10). We focus on testing the symmetry restrictions entailed by microeconomic theory, i.e.,

\begin{align*}
H_{01}: \begin{cases} 
  d_{KL} = d_{LK}, & d_{KM} = d_{MK} \\
  d_{KE} = d_{EK}, & d_{LM} = d_{ML} \\
  d_{LE} = d_{EL}, & d_{EM} = d_{ME}
\end{cases}
\end{align*}

as well as a subset of these constraints:

\begin{align*}
H_{02}: d_{EM} = d_{ME}, & \quad d_{KM} = d_{MK}
\end{align*}
Confirming with the procedures described above, we reconsider the testing problem in the context of the MLR model of which the KLEM system is a restricted form. The individual equations of the latter model include the 32 price ratios \((P_i/P_j)^{1/2}, \ i, j = K, L, E, M\) as regressors. The unrestricted MLE SURE estimates using the data provided in Berndt (1991) on the manufacturing sector of the U.S. economy over the period 1947–1971 are given below (with asymptotic standard errors in parentheses):

\[
\begin{align*}
K/Y &= 0.0263 + 0.0036(P_L/P_K)^{1/2} + 0.0649(P_E/P_K)^{1/2} \\
&\quad - 0.0443(P_M/P_K)^{1/2} + \hat{\epsilon}_K \\
L/Y &= 0.0719 + 0.0517(P_K/P_L)^{1/2} + 0.2200(P_E/P_L)^{1/2} \\
&\quad + 0.0264(P_M/P_L)^{1/2} + \hat{\epsilon}_L \\
E/Y &= 0.0403 - 0.0111(P_K/P_E)^{1/2} - 0.0048(P_L/P_E)^{1/2} \\
&\quad + 0.0150(P_M/P_E)^{1/2} + \hat{\epsilon}_E \\
M/Y &= 0.7401 - 0.0542(P_K/P_M)^{1/2} - 0.1374(P_L/P_M)^{1/2} \\
&\quad + 0.0399(P_E/P_M)^{1/2} + \hat{\epsilon}_M
\end{align*}
\]

For both hypotheses, we computed the GLS and ML-based Wald statistics, Eq. (49) the LR and LR_M criteria as defined in Eqs. (23) and (28) and the QLR statistics, Eq. (48). In the case of the Wald and QLR test, we obtained the asymptotic \(\chi^2\) and LMC \(p\) values using 19 and 99 simulated samples. The exact BMC and MMC \(p\) values were also obtained for the LR criteria. The bounding statistic \(LR_{UL} = n\ln(\Lambda_{UL})\) corresponds to the UL hypothesis that sets all the coefficients of the MLR model (except the intercepts) to specific values. As stated in Section 3, the BMC procedure based on \(LR_M\) yields tighter bounds [see inequality (42)]. Our results are summarized in Tables 6 and 7.

From these results, we see that the symmetry hypothesis \(H_{01}\) is rejected using all asymptotic and exact tests. In the case of \(H_{02}\), all tests against the unconstrained SURE specification are not significant. However, the asymptotic \(\chi^2\) and LMC tests \(LR_M\) are significant at the 5% level. Although the bounds \(p\)-value is larger than 0.05, the MMC test is significant at the 5% level, even with 19 simulated samples. It is worth noting that the QLR and the LR LMC tests yield equivalent decisions for both testing problems. Moreover, all MC tests based on 19 and 99 replications also yield similar decisions.
TABLE 6  Generalized Leontief Factor Demands: Cross-Equation Symmetry Tests

\[ H_{01}^a : \begin{align*} 
  d_{KL} &= d_{LK}, \\
  d_{KM} &= d_{MK}, \\
  d_{KE} &= d_{EK}, \\
  d_{LM} &= d_{ML}, \\
  d_{LE} &= d_{EL}, \\
  d_{EM} &= d_{ME} 
\end{align*} \]

<table>
<thead>
<tr>
<th>Statistic</th>
<th>LR_{ML}</th>
<th>LR</th>
<th>QLR_{(0)}</th>
<th>QLR_{(1)}</th>
<th>QLR_{(2)}</th>
<th>Wald_{GLS}</th>
<th>Wald_{ML}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic p-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Reps(b)</td>
<td>MC p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 BMC</td>
<td>0.05</td>
<td>0.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>99</td>
<td>0.01</td>
<td>0.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19 MMC</td>
<td>0.05</td>
<td>0.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>99</td>
<td>0.01</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19 LMC</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>99</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\(a\)Under \(H_{01}\), \(LR_{ML} \sim \chi^2(42)\) while the other statistics have asymptotic \(\chi^2(6)\) distributions.

The \(LR_{ML}\) statistic tests the symmetry restrictions (6 constraints) jointly with the SURE exclusion restrictions (36 constraints)—a total of 42 restrictions—against the unrestricted MLR model.

\(b\)Replications.

TABLE 7  Generalized Leontief Factor Demands: Partial Cross-Equation Symmetry Tests

\[ H_{02}^a : \begin{align*} 
  d_{EM} &= d_{ME}, \\
  d_{KM} &= d_{MK} 
\end{align*} \]

<table>
<thead>
<tr>
<th>Statistic</th>
<th>LR_{ML}</th>
<th>LR</th>
<th>QLR_{(0)}</th>
<th>QLR_{(1)}</th>
<th>QLR_{(2)}</th>
<th>Wald_{GLS}</th>
<th>Wald_{ML}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic p-value</td>
<td>0.000</td>
<td>0.927</td>
<td>0.927</td>
<td>0.927</td>
<td>0.927</td>
<td>0.937</td>
<td>0.938</td>
</tr>
<tr>
<td>Reps(b)</td>
<td>MC p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 BMC</td>
<td>0.15</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>99</td>
<td>0.17</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19 MMC</td>
<td>0.05</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>99</td>
<td>0.05</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19 LMC</td>
<td>0.05</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>99</td>
<td>0.04</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

\(a\)Under \(H_{02}\), \(LR_{ML} \sim \chi^2(38)\) while the other statistics have asymptotic \(\chi^2(2)\) distributions.

\(\text{LR}_{ML}\) tests a subset of symmetry restrictions (2 constraints) jointly with the SURE exclusion restrictions (36 constraints)—38 restrictions in all—against the unrestricted MLR model.

\(b\)Replications.
6 CONCLUSION

In this chapter we have extended the MLR-based LR test procedure to the SURE framework. We have combined the bounds and Monte Carlo test approaches to provide \( p \)-values for test statistics that can yield provably exact tests in finite samples even for nonlinear hypotheses as well as more reliable large sample tests. The feasibility of the test strategy was also illustrated with an extensive Monte Carlo experiment and an empirical application. We have found that standard asymptotic tests exhibit serious errors in level, particularly in larger systems. In contrast, the various tests we have proposed displayed excellent size and power properties.

ACKNOWLEDGMENTS

This work was supported by the Canadian Network of Centres of Excellence [program on Mathematics of Information Technology and Complex Systems (MITACS)], the Canada Council for the Arts (Killam Fellowship), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds FCAR (Government of Quebec). This chapter is a revised version of an earlier working paper by Dufour and Khalaf (1998). The authors thank Emanuela Cardia, Marcel Dagenais, John Galbraith, Eric Ghysels, James MacKinnon, Christophe Muller, Olivier Torrès, and Michael Veall for useful comments. Earlier versions of the work were presented at the North American Meetings of the Econometric Society, the Third International Conference on Computing and Finance (Hoover Institute, Stanford University), the Annual Meetings of the American Statistical Association, the Canadian Econometric Study group, the Canadian Economic Association, and at Ohio State University (Economics).

REFERENCES


Finding Optimal Penalties for Model Selection in the Linear Regression Model

Maxwell L. King  
Monash University, Melbourne, Australia

Gopal K. Bose  
Queensland Treasury, Brisbane, Queensland, Australia

1 INTRODUCTION

In econometrics and statistics, we often need to make a choice between a number of alternative statistical models. In some cases, economic theory or our understanding of the phenomenon we are trying to model can help us make an appropriate decision. However, there are many situations in which we are forced to ask which model best fits the data. The problem of choosing between a limited range of alternative models using only the available data is known as the model selection problem.

There is a large literature on this problem with a range of methods or strategies being suggested as solutions. A popular method of deciding on a final model using the data has been to use a series of pairwise hypothesis tests. Unfortunately, as noted by Granger et al. (1995), this approach has many limitations. At each step, one model is chosen as the null hypothesis and can be unfairly favored because the probability of wrongly rejecting it is set to a small value like 0.05 or 0.01. This can be
particularly troublesome in situations where the test being used is not particularly powerful and therefore a choice of the null hypothesis model is the most likely outcome. There is also the well known problem of pretest bias (see, e.g., Wallace, 1977; Giles and Giles, 1993). Finally different researchers working on the same model selection problem could easily end up with different final models purely because they performed the tests in a different order or used different significance levels.

The most widely accepted class of model selection procedures is the use of information criteria (IC) based on choosing the model with the largest maximized log-likelihood function minus a penalty that is an increasing function of the number of free parameters included in the model. Examples of these procedures include Akaike’s (1973, 1974) IC (AIC), Schwarz’s (1978) Bayesian IC (BIC), and Hannan and Quinn’s (1979) procedure denoted HQ. Fox (1995) has also demonstrated how a number of other model selection procedures, particularly those developed for choosing from a set of linear regression models, can be thought of in the IC framework. Those of interest in this chapter include the use of Theil’s (1961) \( R^2 \) criterion, Mallows’ (1964) \( C_p \) procedure, denoted MCP, Schmidt’s (1975) generalised cross-validation (GCV) procedure, and Hocking’s (1976) criterion, denoted HOC. As Granger et al. (1995) and others have noted, the IC approach has the advantages that:

1. No particular model is favored because it has been chosen to be a null hypothesis.
2. The order of computation is irrelevant.
3. Pretesting bias is not an issue.
4. If the IC procedure is asymptotically consistent, the correct model is chosen with probability one asymptotically.
5. There is no need to choose an arbitrary level of significance although there is the related issue of which penalty function is appropriate.

The last is a major issue that has featured in the literature. There is currently little agreement about what the form of the penalty function should be. The early literature focused on asymptotic arguments to justify various choices of penalty functions (see, e.g., Akaike, 1973; Schwarz, 1978; Hannan and Quinn, 1979). Since then we have seen a number of Monte Carlo studies of the small sample properties of different model selection procedures, which have shown a number of problems. They suggest that asymptotic properties are no guarantee of acceptable small sample properties. For example, Grose and King (1994), in the context of choosing between first-order autoregressive and first-order moving average disturbances in the linear regression model, have shown that a particular
model can be unfairly favored because of the functional form of its log-likelihood function. They also found that the presence of nuisance parameters can adversely affect the probabilities of correct selection.

There are clear parallels between the hypothesis testing literature and the model selection literature, although it does appear that the latter lags the former. The computer revolution has meant that we can now ask what kind of testing procedures would we like to use rather than what kind of testing procedure is convenient to use (see, e.g., King, 1987). We should be asking similar questions for model selection, in the context of finite samples. We now regularly use simulation to find critical values for hypothesis tests. Can we use similar methods to find penalty values for IC procedures that are in some sense optimal?

In this chapter we investigate a new approach to IC model selection in the context of the classical problem of choosing between different linear regression models. Our approach involves the use of the simulation method to estimate probabilities of correct selection and choosing penalties that optimize these probabilities on average. Maximizing simulated probabilities can be a difficult optimization problem. We use a relatively new optimization algorithm called simulated annealing to overcome any difficulties in this regard.

A similar approach to model selection has been investigated by Kwek (1999) and Kwek and King (1999), in the context of selecting between different conditional heteroscedastic processes. Billah and King (2000a,b) have also considered a similar method for choosing between different time-series processes for linear regression disturbances. Each of these studies has involved choosing between variance–covariance matrix functions with restricted parameter spaces. In this chapter we consider model selection of different mean processes with unrestricted parameter spaces.

The plan of this chapter is as follows. The new approach is outlined in abstract terms in Section 2. The main practical issues of applying this approach to choosing between \( m \) linear regression models are discussed in Section 3. These include the choice of weighting function, the issue of how best to estimate mean average probabilities via simulation and the choice of optimization method. The design of our Monte Carlo study to compare the new approach with some existing IC methods is given in Section 4 and the results of this study are discussed in Section 5. The chapter closes with some concluding remarks.

2 FRAMEWORK FOR OPTIMAL PENALTY SELECTION

Our problem of interest is one of selecting a model from \( m \) alternative models that we will denote by \( M_1, M_2, \ldots, M_m \), for a given set of \( n \)
observations on the dependent variable denoted by the $n \times 1$ vector $y$. Let model $M_j, j = 1, \ldots, m$, be represented by

$$y = f(X_j, \beta_j, u)$$

(1)

where $f$ is a known $n \times 1$ vector function, $X_j$ is an $n \times q_j$ matrix of observations on $q_j$ fixed variables, $\beta_j$ is a $q_j \times 1$ vector of unknown structural parameters and $u$ is an $n \times 1$ vector of random disturbances distributed as $N(0, \sigma_j^2 I_n)$ where $\sigma_j^2$ is a further unknown parameter. Let $\theta_j = (\beta_j, \sigma_j^2)'$ denote the $k_j \times 1$ vector of unknown parameters in this model where $k_j = q_j + 1$. Let $L_j(\theta_j)$ denote the log-likelihood function of model $M_j$ and let $L_j(\hat{\theta}_j)$ denote its maximized value where $\hat{\theta}_j$ is the maximum likelihood estimator (MLE) of $\theta_j$.

Almost all IC-based model selection procedures are based on choosing the model with the largest value of

$$I_j = L_j(\hat{\theta}_j) - p_j$$

(2)

for $j = 1, \ldots, m$, where $p_j$ is the penalty for model $M_j$. Examples of values for $p_j$ include $k_j$ for AIC, $k_j \log(n)/2$ for BIC, $k_j \log[\log(n)]$ for HQ, $-n \log(n - k_j)/2$ for $R^2$, $-n \log((n - k_j)/n)$ for GCV, $-n \log((n - k_j) \times (n - k_j - 1))/2$ for HOC, and $n \log(1 + 2k_j/(n - k_j^2))/2$ for MCP where $k^*$ is the number of free parameters in the smallest model that includes all models under consideration as special cases.

The main question we wish to address is how do we find optimal penalties, $p_1, \ldots, p_m$, for Eq. (2)? We can look to the hypothesis testing literature for some guidance. There we make a choice of test based on maximizing power which is the probability of correctly rejecting the null hypothesis. It therefore seems reasonable to want to choose $p_j$ to somehow maximize the probability of correctly selecting the true model. Unfortunately, these probabilities depend on the true model and the value of $\theta_j$ for this model. Our suggestion of a criterion for optimally choosing $p_1, \ldots, p_m$ is as follows. For any given set of penalty values (i.e., $p_j, j = 1, \ldots, m$) and for each model under consideration, we estimate the average probability of correctly selecting this model when it is indeed the true model. For the same penalty set, we then sum these probabilities of correct selection, thus treating all models equally. The penalties are chosen to maximize this sum.

In order to understand more closely what is involved, let $CS_j$ denote the event of correct selection of $M_j$ when $M_j$ is the true model and let

$$P(CS_j | \theta_j, p_1, \ldots, p_m)$$
denote the probability of correct selection when $M_j$ is the true model with parameter vector $\theta_j$ and $p_1, \ldots, p_m$ are used as the penalties in Eq. (2). Then the average probability of correct selection for the $j$th model with penalties $p_1, \ldots, p_m$ can then be written as

$$APCS_j(p_1, \ldots, p_m) = \int P(CS_j|\theta_j, p_1, \ldots, p_m)g_j(\theta_j) \, d\theta_j$$

(3)

where $g_j(\theta_j)$ is a weighting density function rather like a prior density function used in Bayesian methods. Its purpose is to weight different parameter vector values when calculating the average probability of correct selection. The objective function we wish to choose $p_1, \ldots, p_m$ to maximize is, therefore,

$$MAPCS(p_1, \ldots, p_m) = \frac{1}{m} \sum_{j=1}^{m} APCS_j(p_1, \ldots, p_m)$$

(4)

For our approach to be operational, we need a method of estimating Eq. (3). Given that $g_j(\theta_j)$ is a joint density function, the method of Monte Carlo estimation of Eq. (3) suggests itself. This allows Eq. (3) to be estimated by first taking a large sample of drawings from the distribution implied by $g_j(\theta_j)$. We denote these independent drawings by $\theta_j(i), i = 1, \ldots, r$, where $r$ is the number of drawings. The estimate of Eq. (3) is then obtained by calculating the sample mean:

$$\frac{1}{r} \sum_{i=1}^{r} P(CS_j|\theta_j(i), p_1, \ldots, p_m)$$

(5)

This in turn requires us to estimate

$$P(CS_j|\theta_j(i), p_1, \ldots, p_m)$$

(6)

for each value of $\theta_j(i), p_1, \ldots, p_m, i = 1, \ldots, r$, which can be achieved in the normal manner by a standard Monte Carlo simulation of $N$ replications from the model represented by Eq. (1).

This still leaves a number of issues for resolution before the method can be implemented. These are the choice of $g_j(\theta_j)$ and how we might sample from its implied distribution, the choice of values for $r$ and $N$, and the method for maximizing the estimated value of the $MAPCS(p_1, \ldots, p_m)$. These are addressed in Section 3 in the context of choosing between different linear regression models.
3 SOME PRACTICAL ISSUES

We now turn our attention to the specific problem of choosing between \( m \) regression models of the form:

\[
y = X_j \beta_j + u
\]

where \( y, X_j, \beta_j, \) and \( u \) are defined as in Eq. (1).

3.1 Choice of Weighting Function

In choosing a weighting function for \( \theta_j = (\beta_j, \sigma_j^2)' \), we need to balance the need not to favor unduly a particular part of the \( \theta_j \) space with the need for \( APCS_j(p_1, \ldots, p_m) \) to take a value between 0 and 1 so that it is reasonably sensitive to different values of \( p_1, \ldots, p_m \). For example, if Eq. (3) was always unity because of our choice of \( g_j(\theta_j) \), it would be almost impossible to find unique values of \( p_1, \ldots, p_m \) that maximize Eq. (4) as required. This calls for the use of a publicly respected prior for \( \theta_j \) that can be tightened around the origin should Eq. (4) prove to get too close to unity.

For the standard linear regression model, Eq. (7), Zellner (1971) suggests the use of an inverted gamma distribution for \( \sigma_j^2 \) with scale parameter \( s_2^2 \) and a two-step prior for \( \beta_j \) which involves drawing \( \beta_j \) from the multivariate normal distribution:

\[
N(0, \sigma_j^2 I)
\]

where \( \sigma_j^2 \) itself is a separate drawing from an inverted gamma distribution with scale parameter \( s_2^2 \).

The required \( \theta_j(i) = [\beta_j(i)', \sigma_j^2(i)]', \ i = 1, \ldots, r \), can be generated as follows:

**Step 1:** To generate \( \sigma_j^2(i) \), draw a simple random sample of size \( n - q_j \) denoted \( z_1, \ldots, z_{n-q_j} \) from the \( N(0, 1) \) distribution and then calculate

\[
\sigma_j^2(i) = (n - q_j) s_2^2 \left( \sum_{i=1}^{n-q_j} z_i^2 \right)^{-1}
\]

where \( s_2^2 \) is the predetermined scale parameter which is held constant for all \( r \) drawings.

**Step 2:** Generate \( \sigma_j^2 \) with a new independent sample as outlined in **Step 1** but with \( s_2^2 \) replaced by \( s_b^2 \), another predetermined scale parameter that is held constant for all \( r \) drawings.
Step 3: To generate \( \beta_j(i) \), draw a simple random sample of size \( q_j \) denoted \( w_j = (w_{1j}, \ldots, w_{qj}) \) from the \( N(0, 1) \) distribution and calculate

\[
\beta_j(i) = w_j \hat{\sigma}_j
\]

where \( \hat{\sigma}_j \) is the square root of \( \hat{\sigma}_j^2 \) calculated in Step 2.

Observe that increasing (decreasing) \( s^2_e \) and/or \( s^2_b \) makes it easier (harder) to select the correct model and hence increases (decreases) the probability of correct selection and also of the objective function (4).

3.2 Simulation Experiment to Determine the Best Choice of \( N \) and \( r \) Values

In this subsection, we investigate different choices of values for \( N \), the number of Monte Carlo replications used to estimate the probability of correctly choosing model \( j \) when it is the true model with parameter vector \( \theta_j(i) \) and for \( r \), the number of drawings used to estimate Eq. (3) by Eq. (5). The total number of replications in the whole process of estimating Eq. (3) is, therefore, \( rN \). Our aim is to find the best values of \( r \) and \( N \) for fixed \( rN \). In Subsection 3.2.1, we outline the simulation experiments conducted for this purpose and discuss the results of these experiments in Subsection 3.2.2.

3.2.1 Experimental Design

We conducted the simulation experiment in the context of choosing between the following four linear regression models:

\[
\begin{align*}
M_1: & \quad y_t = \beta_{10} + u_{1t} & u_{1t} & \sim \text{IN}(0, \sigma_1^2) \\
M_2: & \quad y_t = \beta_{20} + x_{1t}\beta_{21} + u_{2t} & u_{2t} & \sim \text{IN}(0, \sigma_2^2) \\
M_3: & \quad y_t = \beta_{30} + x_{2t}\beta_{31} + u_{3t} & u_{3t} & \sim \text{IN}(0, \sigma_3^2) \\
M_4: & \quad y_t = \beta_{40} + x_{1t}\beta_{41} + x_{2t}\beta_{42} + u_{4t} & u_{4t} & \sim \text{IN}(0, \sigma_4^2)
\end{align*}
\]

where \( y_t \) is the \( t \)th observation of the dependent variable, \( x_{it} \) is the \( t \)th observation on the \( i \)th nonconstant regressor, the \( \beta \)'s are the regression coefficients and the \( u_{it} \) are the random disturbances for \( t = 1, \ldots, n \). The following two data sets were used as design matrices for these regression models:

\( X1 \): \( x_{1t} \) is the quarterly Australian retail trade as recorded by the Australian Bureau of Statistics and \( x_{2t} \) is the same series lagged by one quarter.

\( X2 \): \( x_{1t} \) is the real per capita GDP and \( x_{2t} \) is the investment of the \( t \)th country. We used the annual data from Summers and Heston (1991) revised version 5.6 and World Bank world tables.
The experiments were conducted for \( n = 20, 50, \) and 96 for \( X_1 \), and \( n = 20, 50, \) and 100 for \( X_2 \). For each model in turn, parameter values were randomly generated as outlined in Subsection 3.1 with \( s_e^2 = 55 \) and \( s_b^2 = 6 \) for \( X_1 \) and \( s_e^2 = 2.5 \) and \( s_b^2 = 0.002 \) for \( X_2 \). These values were chosen after some experimentation in order to obtain the expected probabilities of correct selection around the middle value of 0.5. As the aim was to find the best choice of \( r \) and \( N \) values for fixed \( rN \), the following 15 combinations of \( N \) and \( r \) were used that give \( rN = 2000 \):

<table>
<thead>
<tr>
<th>( N )</th>
<th>2000</th>
<th>1000</th>
<th>500</th>
<th>250</th>
<th>200</th>
<th>100</th>
<th>50</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>8</th>
<th>5</th>
<th>4</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>250</td>
<td>400</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
</tr>
</tbody>
</table>

For each experiment, the values of the four penalties (one for each model) were held fixed. We used seven sets of penalties, namely those for AIC, BIC, HQ, \( \hat{R}^2 \), GCV, MCP, and HOC noted in Section 2.

Each simulation experiment was repeated with different random numbers 20 times so that the standard deviation of the estimated mean probabilities of correct selection [our estimate of Eq. (3)] can be calculated with each model being the true model in turn. We also computed the average of these standard deviations over the four competing models for each set of penalties with the aim of looking for \( N \) and \( r \) combinations where this average standard deviation (ASD) is lowest (and hence our estimates of Eq. (3), on average, have the least variation).

### 3.2.2 Results

Selected results of the simulation experiments are presented in Tables 1 and 2. From these tables, we see a very clear tendency for the value of ASD to decrease as \( r \) increases, particularly for small \( r \). In almost all cases, the lowest ASD is obtained when the model parameters are generated 2000 times \( (r = 2000) \) and only one replication \( (N = 1) \) is used to estimate \( P(CS_j | \theta_1(t), p_1, \ldots, p_m) \). There are some exceptions to these general trends that may be due in part to sampling error.

We therefore decided to estimate the mathematical relationship between ASD and \( r \). Plots of the data suggested

\[
\log(ASD) = a + b \log(r_i) + u_i
\]

where \( a \) and \( b \) are parameters and \( u_i \) is an error term. Ordinary least squares estimation of these models for different penalty functions and values of \( n \) resulted in strong evidence of positive autocorrelation, suggesting
functional misspecification. Further experimentation led to the model:

$$\log(ASD_i) = a + b \log(r_i) + c(\log(r_i))^2 + u_i$$

(8)

where $u_i \sim IN(0, \sigma^2)$.

We estimated this model for each set of penalties, values of $n$, and set of design matrices using our 15 observations on $ASD_i$ and $r_i$. The resulting estimates, values of $R^2$, and the Durbin–Watson (DW) statistics for $X1$ and $X2$ are given in Tables 3 and 4, respectively.

Our best guess of the relationship between $ASD$ and $r$ is, therefore, given by

$$\log(ASD) = \hat{a} + \hat{b}\log(r) + \hat{c}[\log(r)]^2$$

(9)

We can differentiate Eq. (9) and set the derivative equal to zero in order to find which value of $\log(ASD)$ (and hence $ASD$) is a minimum. This value turns out to be

$$\hat{r} = e^{-\frac{\hat{b}}{2\hat{c}}}$$

(10)

Tables 3 and 4 also give the resulting values of $\hat{r}$ calculated from Eq. (10). All values of $\hat{r}$ for $X1$ are greater than 2000. This is also true for $X2$ when $n = 50$ and 100. Only for $X2$ with $n = 20$ do we get $\hat{r}$ values below 2000. The minimum would seem to be around $r = 1000$ in this particular case.

### 3.2.3 Conclusions

Taken as a whole, these results lead us to use $r = 2000$ and $N = 1$ when estimating Eq. (3). The evidence is particularly compelling for $X1$ for all $n$ values and for $X2$ with $n = 50$ and 100. It certainly gets stronger as $n$ increases.

### 3.3 Issues of Optimization

Our basic problem is one of estimating Eq. (4) and then maximizing it with respect to $p_1, \ldots, p_m$. Without loss of generality, we can set $p_1 = 0$ because, from Eq. (2), decisions about which model is chosen depends on differences between penalties, not on their overall level. Because our objective function is made up of averaged probabilities which in turn have been estimated by the Monte Carlo method, it is essentially a step function. These types of functions are difficult to optimize using standard numerical methods. They have many plateaus that will cause standard optimization techniques great difficulties.

There are two approaches to this problem that we recommend. The first is grid-search which can be very time consuming when $m$ is large. We therefore recommend the use of the simulated annealing (SA) algorithm. From the work of Kirkpatrick et al. (1983), Romeo et al. (1984), White...
<table>
<thead>
<tr>
<th>r</th>
<th>AIC</th>
<th>BIC</th>
<th>GCV</th>
<th>HOC</th>
<th>HQ</th>
<th>$R^2$</th>
<th>MCP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AOM ASD</td>
<td>AOM ASD</td>
<td>AOM ASD</td>
<td>AOM ASD</td>
<td>AOM ASD</td>
<td>AOM ASD</td>
<td>AOM ASD</td>
</tr>
<tr>
<td>n=20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.456 0.1527</td>
<td>0.456 0.1647</td>
<td>0.458 0.1565</td>
<td>0.458 0.1577</td>
<td>0.458 0.1563</td>
<td>0.434 0.1232</td>
<td>0.467 0.1532</td>
</tr>
<tr>
<td>4</td>
<td>0.423 0.0904</td>
<td>0.417 0.0947</td>
<td>0.423 0.0920</td>
<td>0.423 0.0926</td>
<td>0.423 0.0920</td>
<td>0.408 0.0779</td>
<td>0.423 0.0905</td>
</tr>
<tr>
<td>8</td>
<td>0.429 0.0601</td>
<td>0.422 0.0638</td>
<td>0.430 0.0617</td>
<td>0.430 0.0620</td>
<td>0.430 0.0616</td>
<td>0.412 0.0494</td>
<td>0.430 0.0604</td>
</tr>
<tr>
<td>20</td>
<td>0.430 0.0343</td>
<td>0.422 0.0373</td>
<td>0.429 0.0350</td>
<td>0.429 0.0363</td>
<td>0.429 0.0351</td>
<td>0.413 0.0289</td>
<td>0.430 0.0343</td>
</tr>
<tr>
<td>50</td>
<td>0.431 0.0288</td>
<td>0.425 0.0333</td>
<td>0.431 0.0300</td>
<td>0.431 0.0300</td>
<td>0.431 0.0299</td>
<td>0.412 0.0253</td>
<td>0.431 0.0290</td>
</tr>
<tr>
<td>100</td>
<td>0.431 0.0196</td>
<td>0.424 0.0214</td>
<td>0.431 0.0203</td>
<td>0.431 0.0206</td>
<td>0.431 0.0205</td>
<td>0.416 0.0169</td>
<td>0.431 0.0200</td>
</tr>
<tr>
<td>200</td>
<td>0.432 0.0140</td>
<td>0.427 0.0150</td>
<td>0.432 0.0142</td>
<td>0.432 0.0139</td>
<td>0.432 0.0142</td>
<td>0.414 0.0137</td>
<td>0.432 0.0139</td>
</tr>
<tr>
<td>500</td>
<td>0.433 0.0139</td>
<td>0.427 0.0138</td>
<td>0.433 0.0139</td>
<td>0.433 0.0139</td>
<td>0.433 0.0139</td>
<td>0.415 0.0107</td>
<td>0.433 0.0139</td>
</tr>
<tr>
<td>1000</td>
<td>0.431 0.0111</td>
<td>0.426 0.0107</td>
<td>0.431 0.0113</td>
<td>0.431 0.0111</td>
<td>0.431 0.0113</td>
<td>0.413 0.0111</td>
<td>0.431 0.0112</td>
</tr>
<tr>
<td>2000</td>
<td>0.433 0.0111</td>
<td>0.426 0.0099</td>
<td>0.433 0.0107</td>
<td>0.433 0.0105</td>
<td>0.433 0.0105</td>
<td>0.417 0.0127</td>
<td>0.433 0.0108</td>
</tr>
</tbody>
</table>

<p>| n=50 |           |            |            |            |            |            |             |
| 1  | 0.677 0.2022 | 0.704 0.2571 | 0.680 0.2060 | 0.680 0.2068 | 0.698 0.2303 | 0.600 0.1457 | 0.677 0.2023 |
| 4  | 0.665 0.1060 | 0.703 0.1311 | 0.669 0.1077 | 0.669 0.1083 | 0.691 0.1182 | 0.585 0.0785 | 0.665 0.1060 |</p>
<table>
<thead>
<tr>
<th>n</th>
<th>8</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.681</td>
<td>0.676</td>
<td>0.673</td>
<td>0.674</td>
<td>0.672</td>
<td>0.675</td>
<td>0.673</td>
<td>0.670</td>
</tr>
<tr>
<td>20</td>
<td>0.721</td>
<td>0.713</td>
<td>0.709</td>
<td>0.711</td>
<td>0.709</td>
<td>0.713</td>
<td>0.709</td>
<td>0.707</td>
</tr>
<tr>
<td>50</td>
<td>0.685</td>
<td>0.680</td>
<td>0.677</td>
<td>0.678</td>
<td>0.677</td>
<td>0.679</td>
<td>0.677</td>
<td>0.675</td>
</tr>
<tr>
<td>100</td>
<td>0.0790</td>
<td>0.0416</td>
<td>0.0256</td>
<td>0.0130</td>
<td>0.0109</td>
<td>0.0130</td>
<td>0.0109</td>
<td>0.0099</td>
</tr>
<tr>
<td>200</td>
<td>0.685</td>
<td>0.680</td>
<td>0.677</td>
<td>0.678</td>
<td>0.677</td>
<td>0.679</td>
<td>0.677</td>
<td>0.675</td>
</tr>
<tr>
<td>500</td>
<td>0.678</td>
<td>0.673</td>
<td>0.668</td>
<td>0.667</td>
<td>0.667</td>
<td>0.669</td>
<td>0.667</td>
<td>0.665</td>
</tr>
<tr>
<td>1000</td>
<td>0.708</td>
<td>0.698</td>
<td>0.698</td>
<td>0.698</td>
<td>0.698</td>
<td>0.698</td>
<td>0.698</td>
<td>0.696</td>
</tr>
<tr>
<td>2000</td>
<td>0.0870</td>
<td>0.0417</td>
<td>0.0162</td>
<td>0.0130</td>
<td>0.0109</td>
<td>0.0130</td>
<td>0.0109</td>
<td>0.0099</td>
</tr>
</tbody>
</table>

\( n = 96 \)
<table>
<thead>
<tr>
<th>$r$</th>
<th>AIC</th>
<th>ASD</th>
<th>AIC</th>
<th>ASD</th>
<th>AIC</th>
<th>ASD</th>
<th>AIC</th>
<th>ASD</th>
<th>AIC</th>
<th>ASD</th>
<th>AIC</th>
<th>ASD</th>
<th>AIC</th>
<th>ASD</th>
<th>$R^2$</th>
<th>MCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.501</td>
<td>0.0633</td>
<td>0.518</td>
<td>0.0618</td>
<td>0.506</td>
<td>0.0634</td>
<td>0.508</td>
<td>0.0632</td>
<td>0.506</td>
<td>0.0634</td>
<td>0.461</td>
<td>0.0564</td>
<td>0.501</td>
<td>0.0632</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.490</td>
<td>0.0306</td>
<td>0.507</td>
<td>0.0282</td>
<td>0.495</td>
<td>0.0298</td>
<td>0.496</td>
<td>0.0299</td>
<td>0.495</td>
<td>0.0299</td>
<td>0.461</td>
<td>0.0303</td>
<td>0.491</td>
<td>0.0306</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.493</td>
<td>0.0223</td>
<td>0.509</td>
<td>0.0218</td>
<td>0.497</td>
<td>0.0228</td>
<td>0.499</td>
<td>0.0228</td>
<td>0.497</td>
<td>0.0227</td>
<td>0.465</td>
<td>0.0220</td>
<td>0.493</td>
<td>0.0224</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.492</td>
<td>0.0166</td>
<td>0.507</td>
<td>0.0164</td>
<td>0.496</td>
<td>0.0162</td>
<td>0.497</td>
<td>0.0163</td>
<td>0.496</td>
<td>0.0161</td>
<td>0.463</td>
<td>0.0153</td>
<td>0.492</td>
<td>0.0166</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.488</td>
<td>0.0148</td>
<td>0.505</td>
<td>0.0144</td>
<td>0.492</td>
<td>0.0149</td>
<td>0.494</td>
<td>0.0149</td>
<td>0.493</td>
<td>0.0148</td>
<td>0.449</td>
<td>0.0154</td>
<td>0.489</td>
<td>0.0147</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.492</td>
<td>0.0136</td>
<td>0.509</td>
<td>0.0107</td>
<td>0.497</td>
<td>0.0131</td>
<td>0.498</td>
<td>0.0129</td>
<td>0.497</td>
<td>0.0132</td>
<td>0.464</td>
<td>0.0130</td>
<td>0.493</td>
<td>0.0136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.492</td>
<td>0.0105</td>
<td>0.508</td>
<td>0.0099</td>
<td>0.497</td>
<td>0.0103</td>
<td>0.498</td>
<td>0.0099</td>
<td>0.497</td>
<td>0.0104</td>
<td>0.462</td>
<td>0.0114</td>
<td>0.492</td>
<td>0.0102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.492</td>
<td>0.0118</td>
<td>0.509</td>
<td>0.0092</td>
<td>0.496</td>
<td>0.0115</td>
<td>0.498</td>
<td>0.0114</td>
<td>0.496</td>
<td>0.0115</td>
<td>0.464</td>
<td>0.0113</td>
<td>0.492</td>
<td>0.0116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.494</td>
<td>0.0108</td>
<td>0.509</td>
<td>0.0090</td>
<td>0.498</td>
<td>0.0105</td>
<td>0.499</td>
<td>0.0104</td>
<td>0.498</td>
<td>0.0104</td>
<td>0.466</td>
<td>0.0107</td>
<td>0.495</td>
<td>0.0107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.491</td>
<td>0.0102</td>
<td>0.509</td>
<td>0.0083</td>
<td>0.496</td>
<td>0.0092</td>
<td>0.497</td>
<td>0.0088</td>
<td>0.496</td>
<td>0.0093</td>
<td>0.462</td>
<td>0.0094</td>
<td>0.491</td>
<td>0.0101</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2 Average Over Four Models of Mean over 20 Iterations of Estimated Mean Probability of Correct Selection (Denoted AOM) and the Average Standard Deviations for These 20 Iterations Averaged Over Four Models (Denoted ASD) for $X_2$
<table>
<thead>
<tr>
<th></th>
<th>0.554</th>
<th>0.0368</th>
<th>0.567</th>
<th>0.0339</th>
<th>0.556</th>
<th>0.0369</th>
<th>0.565</th>
<th>0.0378</th>
<th>0.504</th>
<th>0.0324</th>
<th>0.564</th>
<th>0.0368</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.551</td>
<td>0.0265</td>
<td>0.564</td>
<td>0.0247</td>
<td>0.552</td>
<td>0.0269</td>
<td>0.553</td>
<td>0.0268</td>
<td>0.563</td>
<td>0.0258</td>
<td>0.503</td>
<td>0.0233</td>
</tr>
<tr>
<td>20</td>
<td>0.550</td>
<td>0.0198</td>
<td>0.563</td>
<td>0.0185</td>
<td>0.552</td>
<td>0.0196</td>
<td>0.553</td>
<td>0.0196</td>
<td>0.562</td>
<td>0.0193</td>
<td>0.500</td>
<td>0.0186</td>
</tr>
<tr>
<td>50</td>
<td>0.547</td>
<td>0.0164</td>
<td>0.560</td>
<td>0.0140</td>
<td>0.549</td>
<td>0.0166</td>
<td>0.549</td>
<td>0.0166</td>
<td>0.559</td>
<td>0.0163</td>
<td>0.499</td>
<td>0.0146</td>
</tr>
<tr>
<td>100</td>
<td>0.552</td>
<td>0.0118</td>
<td>0.567</td>
<td>0.0102</td>
<td>0.554</td>
<td>0.0120</td>
<td>0.554</td>
<td>0.0121</td>
<td>0.565</td>
<td>0.0107</td>
<td>0.504</td>
<td>0.0113</td>
</tr>
<tr>
<td>200</td>
<td>0.549</td>
<td>0.0108</td>
<td>0.561</td>
<td>0.0088</td>
<td>0.551</td>
<td>0.0108</td>
<td>0.551</td>
<td>0.0105</td>
<td>0.561</td>
<td>0.0100</td>
<td>0.502</td>
<td>0.0105</td>
</tr>
<tr>
<td>500</td>
<td>0.547</td>
<td>0.0113</td>
<td>0.563</td>
<td>0.0088</td>
<td>0.549</td>
<td>0.0112</td>
<td>0.550</td>
<td>0.0111</td>
<td>0.560</td>
<td>0.0115</td>
<td>0.499</td>
<td>0.0125</td>
</tr>
<tr>
<td>1000</td>
<td>0.549</td>
<td>0.0099</td>
<td>0.562</td>
<td>0.0067</td>
<td>0.551</td>
<td>0.0097</td>
<td>0.551</td>
<td>0.0096</td>
<td>0.561</td>
<td>0.0082</td>
<td>0.499</td>
<td>0.0108</td>
</tr>
</tbody>
</table>
| 2000 | n = 100
|   | 0.607 | 0.1313 | 0.621 | 0.1449 | 0.608 | 0.1324 | 0.626 | 0.1422 | 0.545 | 0.1036 | 0.607 | 0.1314 |
| 4 | 0.604 | 0.0599 | 0.613 | 0.0626 | 0.604 | 0.0601 | 0.605 | 0.0601 | 0.619 | 0.0634 | 0.542 | 0.0493 | 0.603 | 0.0599 |
| 8 | 0.622 | 0.0412 | 0.634 | 0.0428 | 0.624 | 0.0413 | 0.624 | 0.0413 | 0.641 | 0.0439 | 0.557 | 0.0335 | 0.622 | 0.0412 |
| 20 | 0.621 | 0.0260 | 0.633 | 0.0254 | 0.622 | 0.0264 | 0.622 | 0.0264 | 0.640 | 0.0267 | 0.556 | 0.0221 | 0.621 | 0.0260 |
| 50 | 0.619 | 0.0216 | 0.632 | 0.0195 | 0.620 | 0.0216 | 0.620 | 0.0216 | 0.637 | 0.0207 | 0.554 | 0.0171 | 0.619 | 0.0217 |
| 100 | 0.617 | 0.0148 | 0.632 | 0.0131 | 0.618 | 0.0148 | 0.619 | 0.0148 | 0.637 | 0.0137 | 0.551 | 0.0134 | 0.617 | 0.0148 |
| 200 | 0.619 | 0.0125 | 0.632 | 0.0104 | 0.620 | 0.0125 | 0.621 | 0.0125 | 0.638 | 0.0115 | 0.554 | 0.0121 | 0.619 | 0.0126 |
| 500 | 0.619 | 0.0097 | 0.633 | 0.0081 | 0.620 | 0.0099 | 0.621 | 0.0099 | 0.638 | 0.0096 | 0.555 | 0.0112 | 0.619 | 0.0097 |
| 1000 | 0.619 | 0.0111 | 0.634 | 0.0071 | 0.620 | 0.0113 | 0.620 | 0.0113 | 0.638 | 0.0093 | 0.555 | 0.0120 | 0.619 | 0.0111 |
(1984) and Goffe et al. (1994), it does seem that this algorithm performs well at finding the global maxima in the presence of local maxima and for functions like ours which have plateaus and other ill-behaviour. The algorithm works well because it accepts both uphill and downhill moves in a random but systematic manner thus allowing the algorithm to by-pass local maxima and plateaus.

Much has been written on the SA algorithm. See in particular Corana et al. (1987), Goffe et al. (1994), Billah and King (2000b).

### 4 EXPERIMENTAL DESIGN OF THE SIMULATION STUDY

In order to investigate the small sample properties of our proposed model selection procedure and also to check its feasibility, we conducted a

<table>
<thead>
<tr>
<th>n</th>
<th>Penalty function</th>
<th>$\hat{a}$</th>
<th>$\hat{b}$</th>
<th>$\hat{c}$</th>
<th>$R^2$</th>
<th>DW</th>
<th>$\hat{r}$</th>
<th>$\hat{S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>AIC</td>
<td>-1.800***</td>
<td>-0.594***</td>
<td>0.0301***</td>
<td>0.988***</td>
<td>1.709*</td>
<td>19,286</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>-1.754***</td>
<td>-0.569***</td>
<td>0.0242**</td>
<td>0.987***</td>
<td>1.720*</td>
<td>127,542</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCV</td>
<td>-1.788***</td>
<td>-0.582***</td>
<td>0.0280***</td>
<td>0.989***</td>
<td>1.691*</td>
<td>32,626</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HOC</td>
<td>-1.778***</td>
<td>-0.583***</td>
<td>0.0276***</td>
<td>0.988***</td>
<td>1.571*</td>
<td>38,623</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HQ</td>
<td>-1.788***</td>
<td>-0.582***</td>
<td>0.0278***</td>
<td>0.989***</td>
<td>1.593*</td>
<td>35,159</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBAR</td>
<td>-1.979***</td>
<td>-0.603***</td>
<td>0.0353***</td>
<td>0.979***</td>
<td>1.682*</td>
<td>5,121</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>-1.801***</td>
<td>-0.590***</td>
<td>0.0295***</td>
<td>0.988***</td>
<td>1.668*</td>
<td>22,026</td>
<td></td>
</tr>
</tbody>
</table>

| 50 | AIC              | -1.478***  | -0.565***  | 0.0333***  | 0.987*** | 1.692* | 4,834    |            |
|    | BIC              | -1.280***  | -0.633***  | 0.0249***  | 0.993*** | 1.817* | 331,321  |            |
|    | GCV              | -1.459***  | -0.669***  | 0.0333***  | 0.987*** | 1.583* | 19,829   |            |
|    | HOC              | -1.451***  | -0.661***  | 0.0335***  | 0.987*** | 1.542* | 19,258   |            |
|    | HQ               | -1.369***  | -0.649***  | 0.0297***  | 0.991*** | 1.622* | 55,599   |            |
|    | RBAR             | -1.807***  | -0.630***  | 0.0359***  | 0.989*** | 1.898* | 6,466    |            |
|    | MCP              | -1.478***  | -0.655***  | 0.0332***  | 0.987*** | 1.696* | 19,234   |            |

| 96 | AIC              | -2.264***  | -0.551***  | 0.0293*    | 0.972*** | 2.243* | 12,121   |            |
|    | BIC              | -2.029***  | -0.507***  | 0.0132*    | 0.977*** | 2.159* | 218,991  |            |
|    | GCV              | -2.254***  | -0.551***  | 0.0290*    | 0.972*** | 2.251* | 13,360   |            |
|    | HOC              | -2.254***  | -0.550***  | 0.0288*    | 0.973*** | 2.249* | 14,025   |            |
|    | HQ               | -2.120***  | -0.539***  | 0.0225*    | 0.974*** | 2.211* | 159,178  |            |
|    | RBAR             | -2.573***  | -0.528***  | 0.0338***  | 0.969*** | 2.496* | 2,467    |            |
|    | MCP              | -2.263***  | -0.552***  | 0.0294*    | 0.972*** | 2.237* | 11,941   |            |

***, **, * Denote significant at the 0.1, 1 and 5% levels, respectively, "ns" denotes not significant, and "nd" denotes an inconclusive test.
Model Selection in Linear Regression Model

### Table 4 OLS Parameter Estimates and Diagnostics for the Model Eq. (8) Together with the $\hat{r}$ Value that Minimizes $\log(\text{ASD})$ for $X^2$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Penalty Function</th>
<th>$\hat{a}$</th>
<th>$\hat{b}$</th>
<th>$\hat{c}$</th>
<th>$\hat{R}^2$</th>
<th>DW</th>
<th>$\hat{r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>AIC</td>
<td>-2.681***</td>
<td>-0.568***</td>
<td>0.0434***</td>
<td>0.962***</td>
<td>2.338ns</td>
<td>695</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>-2.682***</td>
<td>-0.587***</td>
<td>0.0417***</td>
<td>0.964***</td>
<td>2.563nd</td>
<td>1,140</td>
</tr>
<tr>
<td></td>
<td>GCV</td>
<td>-2.693***</td>
<td>-0.554***</td>
<td>0.0404***</td>
<td>0.955***</td>
<td>2.300ns</td>
<td>950</td>
</tr>
<tr>
<td></td>
<td>HOC</td>
<td>-2.692***</td>
<td>-0.556***</td>
<td>0.0400***</td>
<td>0.965***</td>
<td>2.328ns</td>
<td>1,043</td>
</tr>
<tr>
<td></td>
<td>HQ</td>
<td>-2.692***</td>
<td>-0.556***</td>
<td>0.0408***</td>
<td>0.957***</td>
<td>2.302ns</td>
<td>910</td>
</tr>
<tr>
<td></td>
<td>RBAR</td>
<td>-2.829***</td>
<td>-0.512***</td>
<td>0.0375***</td>
<td>0.973***</td>
<td>1.843ns</td>
<td>922</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>-2.682***</td>
<td>-0.566***</td>
<td>0.0429***</td>
<td>0.961***</td>
<td>2.418ns</td>
<td>733</td>
</tr>
<tr>
<td>50</td>
<td>AIC</td>
<td>-2.212***</td>
<td>-0.575***</td>
<td>0.0334***</td>
<td>0.981***</td>
<td>2.172ns</td>
<td>5,474</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>-2.236***</td>
<td>-0.583***</td>
<td>0.0292***</td>
<td>0.981***</td>
<td>2.513ns</td>
<td>21,653</td>
</tr>
<tr>
<td></td>
<td>GCV</td>
<td>-2.211***</td>
<td>-0.570***</td>
<td>0.0325***</td>
<td>0.982***</td>
<td>2.205ns</td>
<td>6,433</td>
</tr>
<tr>
<td></td>
<td>HOC</td>
<td>-2.215***</td>
<td>-0.566***</td>
<td>0.0319***</td>
<td>0.982***</td>
<td>2.277ns</td>
<td>7,126</td>
</tr>
<tr>
<td></td>
<td>HQ</td>
<td>-2.186***</td>
<td>-0.582***</td>
<td>0.0321***</td>
<td>0.974***</td>
<td>2.417ns</td>
<td>8,651</td>
</tr>
<tr>
<td></td>
<td>RBAR</td>
<td>-2.400***</td>
<td>-0.562***</td>
<td>0.0369***</td>
<td>0.981***</td>
<td>2.219ns</td>
<td>2,029</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>-2.215***</td>
<td>-0.573***</td>
<td>0.0332***</td>
<td>0.981***</td>
<td>2.196ns</td>
<td>5,594</td>
</tr>
<tr>
<td>100</td>
<td>AIC</td>
<td>-2.035***</td>
<td>-0.491***</td>
<td>0.0192*</td>
<td>0.987***</td>
<td>2.235ns</td>
<td>357,345</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>-1.926***</td>
<td>-0.502***</td>
<td>0.0125*</td>
<td>0.992***</td>
<td>2.537ns</td>
<td>525,573,182</td>
</tr>
<tr>
<td></td>
<td>GCV</td>
<td>-2.023***</td>
<td>-0.499***</td>
<td>0.0202**</td>
<td>0.986***</td>
<td>2.123ns</td>
<td>231,303</td>
</tr>
<tr>
<td></td>
<td>HOC</td>
<td>-2.022***</td>
<td>-0.499***</td>
<td>0.0202*</td>
<td>0.987***</td>
<td>2.129nd</td>
<td>231,303</td>
</tr>
<tr>
<td></td>
<td>HQ</td>
<td>-1.938***</td>
<td>-0.505***</td>
<td>0.0171*</td>
<td>0.989***</td>
<td>2.286ns</td>
<td>2,587,187</td>
</tr>
<tr>
<td></td>
<td>RBAR</td>
<td>-2.243***</td>
<td>-0.528***</td>
<td>0.0295***</td>
<td>0.986***</td>
<td>1.449ns</td>
<td>7,701</td>
</tr>
<tr>
<td></td>
<td>MCP</td>
<td>-2.035***</td>
<td>-0.491***</td>
<td>0.0192*</td>
<td>0.987***</td>
<td>2.244ns</td>
<td>357,345</td>
</tr>
</tbody>
</table>

*, **, *** Denotes significant at the 0.1, 1, and 5% levels, respectively, “ns” denotes not significant, and “nd” denotes an inconclusive test.

A simulation study comparing probabilities of correct selection of eight model selection procedures in a range of settings. The procedures included in the study are the new method outlined above involving the maximization of the estimated value of MAPCS($p_1, \ldots, p_m$) with respect to $p_2, \ldots, p_m$ (when $p_1 = 0$), which we will denote by MEM, and AIC, BIC, HQ, $R^2$, GCV, HOC, and MCP.

The following four linear regression models along with the four models described in Subsection 3.2.1 were used in the study:

- $M_5$: $y_t = \beta_{50} + x_{3t}\beta_{51} + u_{5t}$, $u_{5t} \sim IN(0, \sigma_5^2)$
- $M_6$: $y_t = \beta_{60} + x_{2t}\beta_{62} + x_{3t}\beta_{63} + u_{6t}$, $u_{6t} \sim IN(0, \sigma_6^2)$
- $M_7$: $y_t = \beta_{70} + x_{2t}\beta_{72} + x_{3t}\beta_{73} + u_{7t}$, $u_{7t} \sim IN(0, \sigma_7^2)$
- $M_8$: $y_t = \beta_{80} + x_{1t}\beta_{81} + x_{2t}\beta_{82} + x_{3t}\beta_{83} + u_{8t}$, $u_{8t} \sim IN(0, \sigma_8^2)$
where, as for $M_1 - M_4$, $y_t$ is the $t$th observation of the dependent variable, $x_{it}$ is the $t$th observation on the $i$th nonconstant regressor, the $\beta$'s are regression coefficients, and the $u_{jt}$ are the random disturbance terms for $t = 1, \ldots, n$.

In addition to $X1$ and $X2$ described in Subsection 3.2.1, the following data sets were used to construct design matrices as required:

$X3$: $x_{1t}$ and $x_{2t}$ are randomly and independently generated values from the $N(0, 1)$ distribution.

$X4$: This is an extension of $X1$ where the additional regressor $x_{3t}$ is $x_{1t}$ lagged two quarters.

$X5$: This is an extension of $X2$ where the additional regressor $x_{3t}$ is the price level of consumption in the $t$th country.

$X6$: This is an extension of $X3$ where the additional regressor $x_{3t}$ is a further set of independently generated values from the $N(0, 1)$ distribution.

The simulation experiments involved estimating the probabilities of correct selection when choosing between $M_1$, $M_2$, $M_3$, and $M_4$ in the context of $X1$, $X2$, and $X3$, and choosing between $M_1$, $M_2$, $\ldots$, $M_8$ in the context of $X4$, $X5$, and $X6$. Two thousand replications were used throughout and in the case of $X3$ and $X6$, the randomly generated regressors were generated once and then held constant across the 2000 replications. Sample sizes of $n = 20$, 50, and 100 were used with $X2$, $X3$, $X5$, and $X6$ while $n = 20$, 50, and 96 were used for $X1$ and $X4$. MEM was applied with $s_e^2 = 55$ and $s_b^2 = 6$ for $X1$, $s_e^2 = 0.1$ and $s_b^2 = 0.5$ for $X2$, $s_e^2 = 0.12$ and $s_b^2 = 0.08$ for $X3$ and $X6$, $s_e^2 = 0.1$ and $s_b^2 = 0.15$ for $X4$, and $s_e^2 = 0.35$ and $s_b^2 = 0.02$ for $X5$; $r = 2000$ and $N = 1$ were used to estimate Eq. (4) as described in Sections 2 and 3.

MEM was applied using the SA algorithm. In applying this method of maximizing the estimated value of Eq. (4), we experimented with different sets of starting values, upper and lower boundaries, and temperature reduction values, which are a requirement of the SA algorithm. Because AIC and BIC are so widely used in econometrics and statistics, we used the relative penalties (namely, $0, p_2 - p_1, p_3 - p_1, \ldots, p_m - p_1$) of AIC and BIC as starting values. We also used zeros as starting values. A condition of the SA algorithm is that the starting values are not outside the boundary values of the parameters. The parameters in our case are $p_2, \ldots, p_m$ with $p_1 = 0$ and we selected boundary values with this in mind. When zeros or the relative penalties of AIC were used as the starting values, we used the maximum number of free parameters, $k$, among the competing models, the number of competing models, $m$, as well as $km$ and the arbitrary value 10 as upper boundary
values. For these starting values we used two types of lower boundary values, namely zeros and the negative of each of the upper boundaries. When the relative penalties of BIC were used as starting values, we used $k \log(n)/2$, $m \log(n)/2$, $km$, and the arbitrary value 10 as upper boundaries together with zeros and the negative of $k$, $m$, 10, and $km$.

In all cases, we used four temperature reduction values, namely, 0.1, 0.01, 0.001 and 0.0001. This means that in total, we experimented with 96 combinations of start-up values for the application of SA to the MEM method.

5 THE RESULTS

The estimated average probabilities of choosing the true model (APCS) for each of the models in turn plus the estimated mean average probabilities of correct selection (MAPCS) for each design matrix are given in Tables 5–10. We have also included the standard deviation (SD) of the $m$ individual APCSs that make up the MAPCS. These give some idea of the relative variability in APCS for each model. The smaller the SD is, the less variation in the APCSs, which indicates a greater evenness in the APCS values which does seem desirable. We would prefer model selection procedures that do not unduly favor particular types of models.

Although we performed 96 sets of optimizations using SA for the MEM approach, we only give the results for four cases. These are for the penalties that give the largest MAPCS over all 96 optimizations, as well as the results for the smallest SD. In many cases, the different optimizations often result in exactly the same APCS and MAPCS so we give the results for the modal value of MAPCS together with those for the median MAPCS. As can be seen from the tables, it is not unusual for the largest MAPCS also to be the modal and the median MAPCS. Finally, for $X1$ to $X3$, we give the relative penalties for each of the IC procedures.

5.1 Results when Choosing Between Four Models

We will first discuss the results for the existing IC methods (AIC, BIC, GCV, HOC, HQ, $\bar{R}^2$, and MCP) before turning to the results for the MEM approach.

In each case, BIC has the highest MAPCS of the existing IC procedures, followed by HOC for $n=20$ and HQ for larger sample sizes. BIC has the largest APCS for the minimal model $M_1$ and the lowest APCS for the largest model $M_4$. In contrast, $\bar{R}^2$ has the lowest MAPCS by a very clear margin and has the lowest APCS for $M_1$ and the largest APCS for $M_4$. The remaining existing IC procedures behave in a reasonably similar...
**Table 5** Estimated Average Probabilities (APCS), Mean Average Probabilities (MAPCS), and Standard Deviations (SD) of Average Probabilities of Correct Selection for $X_7$ Together with Relative Penalty Values

<table>
<thead>
<tr>
<th>Criteria</th>
<th>APCS</th>
<th>MAPCS</th>
<th>SD</th>
<th>Relative penalties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_4$</td>
</tr>
<tr>
<td>$n = 20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.6690</td>
<td>0.6545</td>
<td>0.3850</td>
<td>0.2605</td>
</tr>
<tr>
<td>BIC</td>
<td>0.7970</td>
<td>0.6850</td>
<td>0.3485</td>
<td>0.1700</td>
</tr>
<tr>
<td>GCV</td>
<td>0.6985</td>
<td>0.6710</td>
<td>0.3900</td>
<td>0.2295</td>
</tr>
<tr>
<td>HOC</td>
<td>0.7095</td>
<td>0.6745</td>
<td>0.3885</td>
<td>0.2245</td>
</tr>
<tr>
<td>HQ</td>
<td>0.7035</td>
<td>0.6635</td>
<td>0.3835</td>
<td>0.2370</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4505</td>
<td>0.5775</td>
<td>0.4190</td>
<td>0.3785</td>
</tr>
<tr>
<td>MCP</td>
<td>0.6715</td>
<td>0.6565</td>
<td>0.3850</td>
<td>0.2545</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.8015</td>
<td>0.7255</td>
<td>0.3975</td>
<td>0.1095</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.7940</td>
<td>0.6075</td>
<td>0.3635</td>
<td>0.2465</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.8015</td>
<td>0.7255</td>
<td>0.3975</td>
<td>0.1095</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.8015</td>
<td>0.7255</td>
<td>0.3975</td>
<td>0.1095</td>
</tr>
</tbody>
</table>

$n = 50$

<table>
<thead>
<tr>
<th>Criteria</th>
<th>APCS</th>
<th>MAPCS</th>
<th>SD</th>
<th>Relative penalties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_4$</td>
</tr>
<tr>
<td>AIC</td>
<td>0.7190</td>
<td>0.7780</td>
<td>0.6655</td>
<td>0.5925</td>
</tr>
<tr>
<td>BIC</td>
<td>0.9095</td>
<td>0.8830</td>
<td>0.7100</td>
<td>0.4630</td>
</tr>
<tr>
<td>GCV</td>
<td>0.7230</td>
<td>0.7920</td>
<td>0.6715</td>
<td>0.5840</td>
</tr>
<tr>
<td>Model Type</td>
<td>n = 96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>0.7325 0.7940 0.6720 0.5815 0.6960 0.0906 1.0418 1.0418 2.1058</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>0.8125 0.8335 0.6925 0.5315 0.7175 0.1387 1.3641 1.3641 2.7281</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{R}^2$</td>
<td>0.4900 0.6390 0.5800 0.6915 0.6001 0.0864 0.5155 0.5155 1.0418</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>0.7190 0.7785 0.6660 0.5915 0.6887 0.0795 1.0009 1.0009 2.0035</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.9745 0.8650 0.7185 0.4210 0.7448 0.2400 3.3296 2.8476 5.2849</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.8690 0.8165 0.6840 0.5460 0.7289 0.1446 1.7732 1.6949 2.9996</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.9745 0.8650 0.7185 0.4210 0.7448 0.2400 3.3296 2.8476 5.2849</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.9745 0.8650 0.7185 0.4210 0.7448 0.2400 3.3296 2.8476 5.2849</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.7415 0.8155 0.8035 0.8830 0.8109 0.0580 1.0000 1.0000 2.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.9395 0.9445 0.9160 0.8255 0.9063 0.0552 2.2822 2.2822 4.5643</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.7455 0.8195 0.8080 0.8815 0.8136 0.0557 1.0159 1.0159 2.0426</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>0.7480 0.8205 0.8085 0.8815 0.8146 0.0547 1.0213 1.0213 2.0536</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>0.8665 0.8905 0.8740 0.8595 0.8726 0.0133 1.5183 1.5183 3.0366</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{R}^2$</td>
<td>0.5290 0.6600 0.6610 0.9155 0.6914 0.1618 0.5079 0.5079 1.0213</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>0.7415 0.8155 0.8035 0.8830 0.8109 0.0580 1.0003 1.0003 2.0009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.9940 0.9530 0.9550 0.7920 0.9235 0.0897 4.7801 4.1070 7.6732</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.8650 0.8805 0.8820 0.8600 0.8719 0.0110 1.5598 1.4537 2.9997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.9940 0.9530 0.9550 0.7920 0.9235 0.0897 4.7801 4.1070 7.6732</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.9940 0.9530 0.9550 0.7920 0.9235 0.0897 4.7801 4.1070 7.6732</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criteria</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_4$</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>$n = 20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.6800</td>
<td>0.8060</td>
<td>0.1580</td>
<td>0.2220</td>
</tr>
<tr>
<td>BIC</td>
<td>0.8225</td>
<td>0.8815</td>
<td>0.1150</td>
<td>0.1345</td>
</tr>
<tr>
<td>GCV</td>
<td>0.7125</td>
<td>0.8306</td>
<td>0.1495</td>
<td>0.1865</td>
</tr>
<tr>
<td>HOE</td>
<td>0.7195</td>
<td>0.8396</td>
<td>0.1450</td>
<td>0.1820</td>
</tr>
<tr>
<td>HQ</td>
<td>0.7135</td>
<td>0.8205</td>
<td>0.1455</td>
<td>0.1960</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4710</td>
<td>0.6790</td>
<td>0.2195</td>
<td>0.3695</td>
</tr>
<tr>
<td>MCP</td>
<td>0.6810</td>
<td>0.8080</td>
<td>0.1565</td>
<td>0.2165</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.6245</td>
<td>0.5435</td>
<td>0.4650</td>
<td>0.5080</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.5535</td>
<td>0.5440</td>
<td>0.4845</td>
<td>0.5080</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.6255</td>
<td>0.5435</td>
<td>0.4635</td>
<td>0.5080</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.6490</td>
<td>0.5460</td>
<td>0.4265</td>
<td>0.5055</td>
</tr>
<tr>
<td>$n = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.7560</td>
<td>0.8230</td>
<td>0.3165</td>
<td>0.2080</td>
</tr>
<tr>
<td>BIC</td>
<td>0.9260</td>
<td>0.9450</td>
<td>0.2840</td>
<td>0.0800</td>
</tr>
<tr>
<td>GCV</td>
<td>0.7630</td>
<td>0.8320</td>
<td>0.3155</td>
<td>0.2010</td>
</tr>
<tr>
<td>Metric</td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
<td>Value 4</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>HOC</td>
<td>0.7670</td>
<td>0.8340</td>
<td>0.3150</td>
<td>0.1990</td>
</tr>
<tr>
<td>HQ</td>
<td>0.8450</td>
<td>0.8865</td>
<td>0.3085</td>
<td>0.1430</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.5145</td>
<td>0.6660</td>
<td>0.3070</td>
<td>0.3660</td>
</tr>
<tr>
<td>MCP</td>
<td>0.7560</td>
<td>0.8240</td>
<td>0.3165</td>
<td>0.2075</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.8770</td>
<td>0.5170</td>
<td>0.5935</td>
<td>0.5285</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.7765</td>
<td>0.5180</td>
<td>0.6275</td>
<td>0.5275</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.8770</td>
<td>0.5170</td>
<td>0.5935</td>
<td>0.5285</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.8745</td>
<td>0.5170</td>
<td>0.5920</td>
<td>0.5285</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
<th>Value 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>0.7615</td>
<td>0.8540</td>
<td>0.5215</td>
<td>0.4040</td>
<td>0.6330</td>
<td>0.2052</td>
<td>1.0000</td>
<td>1.0000</td>
<td>2.0000</td>
</tr>
<tr>
<td>BIC</td>
<td>0.9435</td>
<td>0.9540</td>
<td>0.5190</td>
<td>0.2095</td>
<td>0.6590</td>
<td>0.3631</td>
<td>2.3026</td>
<td>2.3026</td>
<td>4.6052</td>
</tr>
<tr>
<td>GCV</td>
<td>0.7646</td>
<td>0.8520</td>
<td>0.5225</td>
<td>0.3970</td>
<td>0.6340</td>
<td>0.2107</td>
<td>1.0152</td>
<td>1.0152</td>
<td>2.0409</td>
</tr>
<tr>
<td>HOC</td>
<td>0.7665</td>
<td>0.8520</td>
<td>0.5220</td>
<td>0.3970</td>
<td>0.6344</td>
<td>0.2112</td>
<td>1.0204</td>
<td>1.0204</td>
<td>2.0414</td>
</tr>
<tr>
<td>HQ</td>
<td>0.8745</td>
<td>0.5170</td>
<td>0.5305</td>
<td>0.2990</td>
<td>0.6552</td>
<td>0.2539</td>
<td>1.5272</td>
<td>1.5272</td>
<td>3.0544</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.5380</td>
<td>0.6935</td>
<td>0.4825</td>
<td>0.5435</td>
<td>0.5644</td>
<td>0.0904</td>
<td>0.5076</td>
<td>0.5076</td>
<td>1.0204</td>
</tr>
<tr>
<td>MCP</td>
<td>0.7615</td>
<td>0.8450</td>
<td>0.5215</td>
<td>0.4035</td>
<td>0.6329</td>
<td>0.2054</td>
<td>1.0002</td>
<td>1.0002</td>
<td>2.0009</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.9340</td>
<td>0.7990</td>
<td>0.7380</td>
<td>0.4635</td>
<td>0.7336</td>
<td>0.1978</td>
<td>15.2409</td>
<td>1.7396</td>
<td>16.0419</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.8450</td>
<td>0.7530</td>
<td>0.7405</td>
<td>0.5040</td>
<td>0.7106</td>
<td>0.1454</td>
<td>2.3308</td>
<td>1.1201</td>
<td>2.9778</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.9340</td>
<td>0.7990</td>
<td>0.7375</td>
<td>0.4635</td>
<td>0.7335</td>
<td>0.1978</td>
<td>7.4871</td>
<td>1.7373</td>
<td>8.2880</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.9340</td>
<td>0.7990</td>
<td>0.7355</td>
<td>0.4635</td>
<td>0.7330</td>
<td>0.1978</td>
<td>6.1524</td>
<td>1.7391</td>
<td>6.9534</td>
</tr>
<tr>
<td>Criteria</td>
<td>APCS</td>
<td></td>
<td></td>
<td></td>
<td>MAPCS</td>
<td>SD</td>
<td>p₂</td>
<td>p₃</td>
<td>p₄</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M₁</td>
<td>M₂</td>
<td>M₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>0.6355</td>
<td>0.5995</td>
<td>0.5725</td>
<td>0.5085</td>
<td>0.5790</td>
<td>0.0536</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td>0.7815</td>
<td>0.6210</td>
<td>0.5750</td>
<td>0.4205</td>
<td>0.5995</td>
<td>0.1486</td>
<td>1.4979</td>
<td>1.4979</td>
</tr>
<tr>
<td>GCV</td>
<td></td>
<td>0.6650</td>
<td>0.6155</td>
<td>0.5815</td>
<td>0.4810</td>
<td>0.5857</td>
<td>0.0778</td>
<td>1.0813</td>
<td>1.0813</td>
</tr>
<tr>
<td>HOC</td>
<td></td>
<td>0.6755</td>
<td>0.6165</td>
<td>0.5825</td>
<td>0.4740</td>
<td>0.5871</td>
<td>0.0846</td>
<td>1.1123</td>
<td>1.1123</td>
</tr>
<tr>
<td>HQ</td>
<td></td>
<td>0.6695</td>
<td>0.6090</td>
<td>0.5735</td>
<td>0.4900</td>
<td>0.5865</td>
<td>0.0750</td>
<td>1.0972</td>
<td>1.0972</td>
</tr>
<tr>
<td>$\hat{R}^2$</td>
<td></td>
<td>0.4305</td>
<td>0.5365</td>
<td>0.5320</td>
<td>0.6100</td>
<td>0.5272</td>
<td>0.0737</td>
<td>0.5407</td>
<td>0.5407</td>
</tr>
<tr>
<td>MCP</td>
<td></td>
<td>0.6400</td>
<td>0.6025</td>
<td>0.5750</td>
<td>0.5075</td>
<td>0.5813</td>
<td>0.0559</td>
<td>1.0059</td>
<td>1.0059</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td></td>
<td>0.8740</td>
<td>0.5745</td>
<td>0.5590</td>
<td>0.4110</td>
<td>0.6046</td>
<td>0.1941</td>
<td>2.1992</td>
<td>1.8608</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td></td>
<td>0.8000</td>
<td>0.6100</td>
<td>0.5330</td>
<td>0.4690</td>
<td>0.6005</td>
<td>0.1466</td>
<td>1.5052</td>
<td>1.6540</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td></td>
<td>0.8740</td>
<td>0.5745</td>
<td>0.5590</td>
<td>0.4110</td>
<td>0.6046</td>
<td>0.1941</td>
<td>2.1992</td>
<td>1.8608</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td></td>
<td>0.8740</td>
<td>0.5745</td>
<td>0.5590</td>
<td>0.4110</td>
<td>0.6046</td>
<td>0.1941</td>
<td>2.1992</td>
<td>1.8608</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>0.7050</td>
<td>0.7140</td>
<td>0.6610</td>
<td>0.6610</td>
<td>0.6853</td>
<td>0.0282</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td>0.8985</td>
<td>0.7660</td>
<td>0.6980</td>
<td>0.5605</td>
<td>0.7308</td>
<td>0.1408</td>
<td>1.9560</td>
<td>1.9560</td>
</tr>
<tr>
<td>GCV</td>
<td></td>
<td>0.7140</td>
<td>0.7235</td>
<td>0.6720</td>
<td>0.6565</td>
<td>0.6915</td>
<td>0.0323</td>
<td>1.0310</td>
<td>1.0310</td>
</tr>
<tr>
<td>Parameter</td>
<td>Mean 1</td>
<td>Mean 2</td>
<td>Mean 3</td>
<td>Mean 4</td>
<td>Mean 5</td>
<td>Mean 6</td>
<td>Mean 7</td>
<td>Mean 8</td>
<td>Mean 9</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>HOC</td>
<td>0.7165</td>
<td>0.7265</td>
<td>0.6740</td>
<td>0.6545</td>
<td>0.6929</td>
<td>0.0342</td>
<td>1.0418</td>
<td>1.0418</td>
<td>2.1058</td>
</tr>
<tr>
<td>HQ</td>
<td>0.7945</td>
<td>0.7460</td>
<td>0.6875</td>
<td>0.6140</td>
<td>0.7105</td>
<td>0.0778</td>
<td>1.3641</td>
<td>1.3641</td>
<td>2.7281</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4735</td>
<td>0.6145</td>
<td>0.5735</td>
<td>0.7455</td>
<td>0.6018</td>
<td>0.1127</td>
<td>0.5155</td>
<td>0.5155</td>
<td>1.0418</td>
</tr>
<tr>
<td>MCP</td>
<td>0.7050</td>
<td>0.7145</td>
<td>0.6615</td>
<td>0.6605</td>
<td>0.6854</td>
<td>0.0284</td>
<td>1.0009</td>
<td>1.0009</td>
<td>2.0035</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.9060</td>
<td>0.7665</td>
<td>0.7035</td>
<td>0.5560</td>
<td>0.7330</td>
<td>0.1452</td>
<td>2.0229</td>
<td>1.9573</td>
<td>4.0287</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.8005</td>
<td>0.7780</td>
<td>0.7010</td>
<td>0.5930</td>
<td>0.7181</td>
<td>0.0937</td>
<td>1.3051</td>
<td>1.4416</td>
<td>2.9445</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.9060</td>
<td>0.7665</td>
<td>0.7035</td>
<td>0.5560</td>
<td>0.7330</td>
<td>0.1452</td>
<td>2.0229</td>
<td>1.9573</td>
<td>4.0287</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.9060</td>
<td>0.7665</td>
<td>0.7035</td>
<td>0.5560</td>
<td>0.7330</td>
<td>0.1452</td>
<td>2.0229</td>
<td>1.9573</td>
<td>4.0287</td>
</tr>
<tr>
<td>AIC</td>
<td>0.7025</td>
<td>0.7665</td>
<td>0.7125</td>
<td>0.7515</td>
<td>0.7332</td>
<td>0.0306</td>
<td>1.0000</td>
<td>1.0000</td>
<td>2.0000</td>
</tr>
<tr>
<td>BIC</td>
<td>0.9315</td>
<td>0.8325</td>
<td>0.7580</td>
<td>0.6375</td>
<td>0.7899</td>
<td>0.1240</td>
<td>2.3026</td>
<td>2.3026</td>
<td>4.6052</td>
</tr>
<tr>
<td>GCV</td>
<td>0.7090</td>
<td>0.7710</td>
<td>0.7170</td>
<td>0.7475</td>
<td>0.7361</td>
<td>0.0286</td>
<td>1.0152</td>
<td>1.0152</td>
<td>2.0409</td>
</tr>
<tr>
<td>HOC</td>
<td>0.7090</td>
<td>0.7710</td>
<td>0.7175</td>
<td>0.7470</td>
<td>0.7361</td>
<td>0.0284</td>
<td>1.0204</td>
<td>1.0204</td>
<td>2.0514</td>
</tr>
<tr>
<td>HQ</td>
<td>0.8450</td>
<td>0.8095</td>
<td>0.7605</td>
<td>0.6980</td>
<td>0.7782</td>
<td>0.0637</td>
<td>1.5272</td>
<td>1.5272</td>
<td>3.0544</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4575</td>
<td>0.6440</td>
<td>0.6055</td>
<td>0.8275</td>
<td>0.6336</td>
<td>0.1522</td>
<td>0.5076</td>
<td>0.5076</td>
<td>1.0204</td>
</tr>
<tr>
<td>MCP</td>
<td>0.7025</td>
<td>0.7665</td>
<td>0.7125</td>
<td>0.7515</td>
<td>0.7332</td>
<td>0.0306</td>
<td>1.0002</td>
<td>1.0002</td>
<td>2.0009</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.8905</td>
<td>0.8330</td>
<td>0.8166</td>
<td>0.6305</td>
<td>0.7926</td>
<td>0.1126</td>
<td>2.1758</td>
<td>1.5387</td>
<td>4.2676</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.8485</td>
<td>0.7945</td>
<td>0.7700</td>
<td>0.7065</td>
<td>0.7799</td>
<td>0.0589</td>
<td>1.6293</td>
<td>1.4306</td>
<td>2.9826</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.8905</td>
<td>0.8330</td>
<td>0.8160</td>
<td>0.6305</td>
<td>0.7925</td>
<td>0.1126</td>
<td>2.1811</td>
<td>1.5461</td>
<td>4.2745</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.8905</td>
<td>0.8330</td>
<td>0.8160</td>
<td>0.6305</td>
<td>0.7925</td>
<td>0.1126</td>
<td>2.1811</td>
<td>1.5461</td>
<td>4.2745</td>
</tr>
<tr>
<td>Criteria</td>
<td>$M_1$</td>
<td>$M_2$</td>
<td>$M_3$</td>
<td>$M_4$</td>
<td>$M_5$</td>
<td>$M_6$</td>
<td>$M_7$</td>
<td>$M_8$</td>
<td>MAPCS</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>$n = 20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.5320</td>
<td>0.3800</td>
<td>0.3635</td>
<td>0.2760</td>
<td>0.3470</td>
<td>0.2625</td>
<td>0.2560</td>
<td>0.2585</td>
<td>0.3344</td>
</tr>
<tr>
<td>BIC</td>
<td>0.7105</td>
<td>0.4060</td>
<td>0.3755</td>
<td>0.2280</td>
<td>0.3655</td>
<td>0.2185</td>
<td>0.2160</td>
<td>0.1905</td>
<td>0.3388</td>
</tr>
<tr>
<td>GCV</td>
<td>0.5730</td>
<td>0.3955</td>
<td>0.3745</td>
<td>0.2670</td>
<td>0.3625</td>
<td>0.2590</td>
<td>0.2475</td>
<td>0.2210</td>
<td>0.3375</td>
</tr>
<tr>
<td>HOC</td>
<td>0.5880</td>
<td>0.4000</td>
<td>0.3760</td>
<td>0.2655</td>
<td>0.3635</td>
<td>0.2570</td>
<td>0.2435</td>
<td>0.2160</td>
<td>0.3387</td>
</tr>
<tr>
<td>HQ</td>
<td>0.5760</td>
<td>0.3965</td>
<td>0.3630</td>
<td>0.2675</td>
<td>0.3555</td>
<td>0.2585</td>
<td>0.2470</td>
<td>0.2380</td>
<td>0.3365</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.2990</td>
<td>0.2990</td>
<td>0.3220</td>
<td>0.2355</td>
<td>0.2910</td>
<td>0.3010</td>
<td>0.2905</td>
<td>0.3590</td>
<td>0.3096</td>
</tr>
<tr>
<td>MCP</td>
<td>0.5325</td>
<td>0.3820</td>
<td>0.3640</td>
<td>0.2740</td>
<td>0.3460</td>
<td>0.2615</td>
<td>0.2525</td>
<td>0.2505</td>
<td>0.3329</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.6320</td>
<td>0.4095</td>
<td>0.4280</td>
<td>0.2650</td>
<td>0.3425</td>
<td>0.2550</td>
<td>0.1870</td>
<td>0.2305</td>
<td>0.3437</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.5695</td>
<td>0.3860</td>
<td>0.4620</td>
<td>0.2410</td>
<td>0.3745</td>
<td>0.2810</td>
<td>0.2320</td>
<td>0.1955</td>
<td>0.3426</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.6255</td>
<td>0.4100</td>
<td>0.4115</td>
<td>0.2685</td>
<td>0.3345</td>
<td>0.2555</td>
<td>0.2465</td>
<td>0.1950</td>
<td>0.3434</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.6245</td>
<td>0.4170</td>
<td>0.4170</td>
<td>0.2650</td>
<td>0.3395</td>
<td>0.2555</td>
<td>0.2345</td>
<td>0.1930</td>
<td>0.3433</td>
</tr>
<tr>
<td>$n = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.6240</td>
<td>0.5700</td>
<td>0.5905</td>
<td>0.5540</td>
<td>0.5946</td>
<td>0.5290</td>
<td>0.5635</td>
<td>0.5725</td>
<td>0.5747</td>
</tr>
<tr>
<td>BIC</td>
<td>0.8770</td>
<td>0.7125</td>
<td>0.7325</td>
<td>0.5280</td>
<td>0.7300</td>
<td>0.5115</td>
<td>0.5400</td>
<td>0.4565</td>
<td>0.6360</td>
</tr>
<tr>
<td>GCV</td>
<td>0.6450</td>
<td>0.5870</td>
<td>0.6040</td>
<td>0.5530</td>
<td>0.6085</td>
<td>0.5305</td>
<td>0.5675</td>
<td>0.5580</td>
<td>0.5817</td>
</tr>
<tr>
<td>Method</td>
<td>HOC</td>
<td>HQ</td>
<td>R²</td>
<td>MCP</td>
<td>MEM with largest MAPCS</td>
<td>MEM with smallest SD</td>
<td>MEM with modal MAPCS</td>
<td>MEM with median MAPCS</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-------------------------</td>
<td>----------------------</td>
<td>----------------------</td>
<td>----------------------</td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>0.6505</td>
<td>0.5900</td>
<td>0.5630</td>
<td>0.6130</td>
<td>0.5315</td>
<td>0.5685</td>
<td>0.5575</td>
<td>0.5836</td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>0.7530</td>
<td>0.6470</td>
<td>0.6640</td>
<td>0.6765</td>
<td>0.5300</td>
<td>0.5690</td>
<td>0.5200</td>
<td>0.6141</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.3620</td>
<td>0.4010</td>
<td>0.4220</td>
<td>0.5010</td>
<td>0.4090</td>
<td>0.4765</td>
<td>0.5230</td>
<td>0.6660</td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>0.6255</td>
<td>0.5710</td>
<td>0.5910</td>
<td>0.5640</td>
<td>0.5955</td>
<td>0.5295</td>
<td>0.5635</td>
<td>0.5720</td>
<td></td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.9255</td>
<td>0.6885</td>
<td>0.7245</td>
<td>0.5625</td>
<td>0.7180</td>
<td>0.5330</td>
<td>0.5300</td>
<td>0.4530</td>
<td></td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.7720</td>
<td>0.6945</td>
<td>0.6485</td>
<td>0.5685</td>
<td>0.6845</td>
<td>0.4940</td>
<td>0.5335</td>
<td>0.5475</td>
<td></td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.9135</td>
<td>0.6940</td>
<td>0.7250</td>
<td>0.5665</td>
<td>0.7275</td>
<td>0.5320</td>
<td>0.5160</td>
<td>0.4680</td>
<td></td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.9135</td>
<td>0.6940</td>
<td>0.7250</td>
<td>0.5665</td>
<td>0.7275</td>
<td>0.5320</td>
<td>0.5160</td>
<td>0.4680</td>
<td></td>
</tr>
</tbody>
</table>

n = 96

<table>
<thead>
<tr>
<th>Method</th>
<th>AIC</th>
<th>BIC</th>
<th>GCV</th>
<th>HOC</th>
<th>HQ</th>
<th>R²</th>
<th>MCP</th>
<th>MEM with largest MAPCS</th>
<th>MEM with smallest SD</th>
<th>MEM with modal MAPCS</th>
<th>MEM with median MAPCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>0.6895</td>
<td>0.6565</td>
<td>0.6860</td>
<td>0.7480</td>
<td>0.7490</td>
<td>0.7210</td>
<td>0.8735</td>
<td>0.7233</td>
<td>0.0702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.9525</td>
<td>0.8850</td>
<td>0.8905</td>
<td>0.8340</td>
<td>0.8340</td>
<td>0.8030</td>
<td>0.8055</td>
<td>0.8135</td>
<td>0.0531</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.6990</td>
<td>0.6680</td>
<td>0.6935</td>
<td>0.7535</td>
<td>0.7535</td>
<td>0.7535</td>
<td>0.7235</td>
<td>0.7291</td>
<td>0.0657</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>0.7025</td>
<td>0.6885</td>
<td>0.6945</td>
<td>0.7555</td>
<td>0.7555</td>
<td>0.7546</td>
<td>0.7250</td>
<td>0.7290</td>
<td>0.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>0.8525</td>
<td>0.7895</td>
<td>0.8130</td>
<td>0.9065</td>
<td>0.7945</td>
<td>0.7845</td>
<td>0.7780</td>
<td>0.8460</td>
<td>0.0278</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.3975</td>
<td>0.4445</td>
<td>0.4600</td>
<td>0.6270</td>
<td>0.6485</td>
<td>0.6345</td>
<td>0.6070</td>
<td>0.9040</td>
<td>0.5654</td>
<td>0.1658</td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>0.6895</td>
<td>0.6575</td>
<td>0.6860</td>
<td>0.7480</td>
<td>0.7490</td>
<td>0.7210</td>
<td>0.8735</td>
<td>0.7235</td>
<td>0.0700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.9970</td>
<td>0.9230</td>
<td>0.9385</td>
<td>0.9380</td>
<td>0.9300</td>
<td>0.8150</td>
<td>0.7870</td>
<td>0.7690</td>
<td>0.8748</td>
<td>0.0831</td>
<td></td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.7145</td>
<td>0.8480</td>
<td>0.8470</td>
<td>0.7920</td>
<td>0.8285</td>
<td>0.7485</td>
<td>0.7720</td>
<td>0.8475</td>
<td>0.7995</td>
<td>0.0615</td>
<td></td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.9970</td>
<td>0.9230</td>
<td>0.9385</td>
<td>0.9380</td>
<td>0.9300</td>
<td>0.8150</td>
<td>0.7870</td>
<td>0.7690</td>
<td>0.8748</td>
<td>0.0831</td>
<td></td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.9970</td>
<td>0.9240</td>
<td>0.9350</td>
<td>0.8390</td>
<td>0.9310</td>
<td>0.8140</td>
<td>0.7825</td>
<td>0.7745</td>
<td>0.8746</td>
<td>0.0825</td>
<td></td>
</tr>
<tr>
<td>Criteria</td>
<td>MAPCS</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.5445</td>
<td>0.4257</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.7240</td>
<td>0.4511</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.5865</td>
<td>0.4359</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>0.6000</td>
<td>0.4394</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>0.5885</td>
<td>0.4331</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.3175</td>
<td>0.4265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>0.5465</td>
<td>0.4265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.6815</td>
<td>0.4939</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.5190</td>
<td>0.4734</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.6830</td>
<td>0.4933</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.6770</td>
<td>0.4922</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\hat{\alpha}$ = \frac{\hat{\beta}}{\hat{\gamma}}$
<table>
<thead>
<tr>
<th>Model Type</th>
<th>AIC</th>
<th>BIC</th>
<th>GCV</th>
<th>HOC</th>
<th>HQ</th>
<th>R²</th>
<th>MCP</th>
<th>MEM with largest MAPCS</th>
<th>MEM with smallest SD</th>
<th>MEM with modal MAPCS</th>
<th>MEM with median MAPCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOC</td>
<td>0.5965</td>
<td>0.7165</td>
<td>0.3675</td>
<td>0.3790</td>
<td>0.6365</td>
<td>0.7600</td>
<td>0.3950</td>
<td>0.4665</td>
<td>0.5384</td>
<td>0.1584</td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>0.7120</td>
<td>0.7950</td>
<td>0.3710</td>
<td>0.3605</td>
<td>0.7015</td>
<td>0.7900</td>
<td>0.3725</td>
<td>0.4045</td>
<td>0.5621</td>
<td>0.2036</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.3215</td>
<td>0.4665</td>
<td>0.2890</td>
<td>0.3965</td>
<td>0.4230</td>
<td>0.6115</td>
<td>0.4105</td>
<td>0.5760</td>
<td>0.4368</td>
<td>0.1124</td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>0.5795</td>
<td>0.6925</td>
<td>0.3695</td>
<td>0.3875</td>
<td>0.6210</td>
<td>0.7486</td>
<td>0.3970</td>
<td>0.4690</td>
<td>0.5318</td>
<td>0.1491</td>
<td></td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.7875</td>
<td>0.7685</td>
<td>0.5530</td>
<td>0.5260</td>
<td>0.7260</td>
<td>0.7060</td>
<td>0.4620</td>
<td>0.4460</td>
<td>0.6218</td>
<td>0.1402</td>
<td></td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.5950</td>
<td>0.6810</td>
<td>0.5380</td>
<td>0.5035</td>
<td>0.6775</td>
<td>0.7095</td>
<td>0.4875</td>
<td>0.4905</td>
<td>0.5836</td>
<td>0.0910</td>
<td></td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.7995</td>
<td>0.7585</td>
<td>0.5560</td>
<td>0.5200</td>
<td>0.7090</td>
<td>0.7165</td>
<td>0.4650</td>
<td>0.4485</td>
<td>0.6211</td>
<td>0.1387</td>
<td></td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.7845</td>
<td>0.7660</td>
<td>0.5505</td>
<td>0.5220</td>
<td>0.7245</td>
<td>0.7095</td>
<td>0.4565</td>
<td>0.4465</td>
<td>0.6200</td>
<td>0.1407</td>
<td></td>
</tr>
</tbody>
</table>

\( n = 100 \)
<table>
<thead>
<tr>
<th>Criteria</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
<th>$M_5$</th>
<th>$M_6$</th>
<th>$M_7$</th>
<th>$M_8$</th>
<th>MAPCS</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.5310</td>
<td>0.3275</td>
<td>0.5425</td>
<td>0.3325</td>
<td>0.5705</td>
<td>0.3515</td>
<td>0.5650</td>
<td>0.3740</td>
<td>0.4493</td>
<td>0.1116</td>
</tr>
<tr>
<td>BIC</td>
<td>0.7065</td>
<td>0.3285</td>
<td>0.6260</td>
<td>0.2885</td>
<td>0.6865</td>
<td>0.3190</td>
<td>0.5770</td>
<td>0.3025</td>
<td>0.4793</td>
<td>0.1858</td>
</tr>
<tr>
<td>GCV</td>
<td>0.5655</td>
<td>0.3445</td>
<td>0.5730</td>
<td>0.3255</td>
<td>0.6125</td>
<td>0.3530</td>
<td>0.5775</td>
<td>0.3385</td>
<td>0.4613</td>
<td>0.1302</td>
</tr>
<tr>
<td>HOC</td>
<td>0.5775</td>
<td>0.3460</td>
<td>0.5835</td>
<td>0.3220</td>
<td>0.6195</td>
<td>0.3525</td>
<td>0.5775</td>
<td>0.3330</td>
<td>0.4639</td>
<td>0.1352</td>
</tr>
<tr>
<td>HQ</td>
<td>0.5690</td>
<td>0.3345</td>
<td>0.5610</td>
<td>0.3280</td>
<td>0.6010</td>
<td>0.3460</td>
<td>0.5660</td>
<td>0.3525</td>
<td>0.4572</td>
<td>0.1258</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.3070</td>
<td>0.2810</td>
<td>0.4020</td>
<td>0.3630</td>
<td>0.3940</td>
<td>0.3840</td>
<td>0.5090</td>
<td>0.4890</td>
<td>0.3899</td>
<td>0.0795</td>
</tr>
<tr>
<td>MCP</td>
<td>0.5325</td>
<td>0.3300</td>
<td>0.5455</td>
<td>0.3335</td>
<td>0.5750</td>
<td>0.3550</td>
<td>0.5680</td>
<td>0.3675</td>
<td>0.4509</td>
<td>0.1129</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.7615</td>
<td>0.4425</td>
<td>0.6540</td>
<td>0.3435</td>
<td>0.6940</td>
<td>0.3455</td>
<td>0.4360</td>
<td>0.3675</td>
<td>0.5056</td>
<td>0.1702</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.5170</td>
<td>0.5160</td>
<td>0.6735</td>
<td>0.3215</td>
<td>0.6520</td>
<td>0.3730</td>
<td>0.3760</td>
<td>0.4835</td>
<td>0.4891</td>
<td>0.1290</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.7520</td>
<td>0.4460</td>
<td>0.6540</td>
<td>0.3450</td>
<td>0.6855</td>
<td>0.3595</td>
<td>0.4355</td>
<td>0.3665</td>
<td>0.5054</td>
<td>0.1648</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.7580</td>
<td>0.4515</td>
<td>0.6545</td>
<td>0.3440</td>
<td>0.6625</td>
<td>0.3720</td>
<td>0.4360</td>
<td>0.3605</td>
<td>0.5047</td>
<td>0.1619</td>
</tr>
<tr>
<td></td>
<td>$n = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.5815</td>
<td>0.4510</td>
<td>0.6105</td>
<td>0.4730</td>
<td>0.5990</td>
<td>0.4900</td>
<td>0.6810</td>
<td>0.5505</td>
<td>0.5546</td>
<td>0.0787</td>
</tr>
<tr>
<td>BIC</td>
<td>0.8675</td>
<td>0.4735</td>
<td>0.7660</td>
<td>0.4165</td>
<td>0.7895</td>
<td>0.4375</td>
<td>0.7155</td>
<td>0.4190</td>
<td>0.6106</td>
<td>0.1913</td>
</tr>
<tr>
<td>GCV</td>
<td>0.5940</td>
<td>0.4575</td>
<td>0.6270</td>
<td>0.4785</td>
<td>0.6190</td>
<td>0.4945</td>
<td>0.6905</td>
<td>0.5400</td>
<td>0.5626</td>
<td>0.0828</td>
</tr>
<tr>
<td>Model Type</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td>MAPCS</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>HOC</td>
<td>0.6010</td>
<td>0.4600</td>
<td>0.6295</td>
<td>0.4790</td>
<td>0.6225</td>
<td>0.4925</td>
<td>0.6895</td>
<td>0.5370</td>
<td>0.5639</td>
<td>0.0834</td>
</tr>
<tr>
<td>HQ</td>
<td>0.7165</td>
<td>0.4770</td>
<td>0.6960</td>
<td>0.4605</td>
<td>0.7055</td>
<td>0.4795</td>
<td>0.6985</td>
<td>0.4965</td>
<td>0.5911</td>
<td>0.1213</td>
</tr>
<tr>
<td>$\hat{R}^2$</td>
<td>0.3105</td>
<td>0.3460</td>
<td>0.4265</td>
<td>0.4730</td>
<td>0.4105</td>
<td>0.4705</td>
<td>0.5685</td>
<td>0.6570</td>
<td>0.4679</td>
<td>0.1134</td>
</tr>
<tr>
<td>MCP</td>
<td>0.5815</td>
<td>0.4610</td>
<td>0.6125</td>
<td>0.4730</td>
<td>0.5995</td>
<td>0.4905</td>
<td>0.6810</td>
<td>0.5500</td>
<td>0.5549</td>
<td>0.0789</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.8560</td>
<td>0.5595</td>
<td>0.7680</td>
<td>0.4675</td>
<td>0.7210</td>
<td>0.5220</td>
<td>0.6500</td>
<td>0.4485</td>
<td>0.6241</td>
<td>0.1486</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.6155</td>
<td>0.6325</td>
<td>0.7045</td>
<td>0.4880</td>
<td>0.7010</td>
<td>0.5030</td>
<td>0.5965</td>
<td>0.5635</td>
<td>0.6006</td>
<td>0.0808</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.8225</td>
<td>0.6085</td>
<td>0.7420</td>
<td>0.4665</td>
<td>0.7125</td>
<td>0.5320</td>
<td>0.6575</td>
<td>0.4480</td>
<td>0.6236</td>
<td>0.1347</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.8390</td>
<td>0.5735</td>
<td>0.7660</td>
<td>0.4640</td>
<td>0.7255</td>
<td>0.5195</td>
<td>0.6510</td>
<td>0.4490</td>
<td>0.6234</td>
<td>0.1450</td>
</tr>
<tr>
<td>AIC</td>
<td>0.5860</td>
<td>0.5310</td>
<td>0.6470</td>
<td>0.5246</td>
<td>0.6660</td>
<td>0.5740</td>
<td>0.7245</td>
<td>0.6465</td>
<td>0.6124</td>
<td>0.0700</td>
</tr>
<tr>
<td>BIC</td>
<td>0.8865</td>
<td>0.5875</td>
<td>0.8220</td>
<td>0.4905</td>
<td>0.8540</td>
<td>0.5390</td>
<td>0.7685</td>
<td>0.5065</td>
<td>0.6818</td>
<td>0.1670</td>
</tr>
<tr>
<td>GCV</td>
<td>0.5905</td>
<td>0.5385</td>
<td>0.6615</td>
<td>0.5266</td>
<td>0.6755</td>
<td>0.5770</td>
<td>0.7300</td>
<td>0.6390</td>
<td>0.6161</td>
<td>0.0702</td>
</tr>
<tr>
<td>HOC</td>
<td>0.5925</td>
<td>0.5390</td>
<td>0.6650</td>
<td>0.5266</td>
<td>0.6755</td>
<td>0.5766</td>
<td>0.7315</td>
<td>0.6385</td>
<td>0.6169</td>
<td>0.0707</td>
</tr>
<tr>
<td>HQ</td>
<td>0.7540</td>
<td>0.5855</td>
<td>0.7570</td>
<td>0.5296</td>
<td>0.7785</td>
<td>0.5825</td>
<td>0.7590</td>
<td>0.5840</td>
<td>0.6663</td>
<td>0.1043</td>
</tr>
<tr>
<td>$\hat{R}^2$</td>
<td>0.3105</td>
<td>0.3870</td>
<td>0.4345</td>
<td>0.4880</td>
<td>0.4430</td>
<td>0.5110</td>
<td>0.6160</td>
<td>0.7415</td>
<td>0.4914</td>
<td>0.1351</td>
</tr>
<tr>
<td>MCP</td>
<td>0.5860</td>
<td>0.5310</td>
<td>0.6470</td>
<td>0.5246</td>
<td>0.6670</td>
<td>0.5740</td>
<td>0.7245</td>
<td>0.6465</td>
<td>0.6126</td>
<td>0.0701</td>
</tr>
<tr>
<td>MEM with largest MAPCS</td>
<td>0.8700</td>
<td>0.6725</td>
<td>0.8025</td>
<td>0.5630</td>
<td>0.7955</td>
<td>0.6236</td>
<td>0.7065</td>
<td>0.5500</td>
<td>0.6879</td>
<td>0.1174</td>
</tr>
<tr>
<td>MEM with smallest SD</td>
<td>0.6715</td>
<td>0.6365</td>
<td>0.7400</td>
<td>0.5590</td>
<td>0.7920</td>
<td>0.5820</td>
<td>0.6535</td>
<td>0.6600</td>
<td>0.6181</td>
<td>0.0764</td>
</tr>
<tr>
<td>MEM with modal MAPCS</td>
<td>0.8490</td>
<td>0.6855</td>
<td>0.8055</td>
<td>0.5630</td>
<td>0.8010</td>
<td>0.6296</td>
<td>0.6955</td>
<td>0.5520</td>
<td>0.6976</td>
<td>0.1130</td>
</tr>
<tr>
<td>MEM with median MAPCS</td>
<td>0.8705</td>
<td>0.6725</td>
<td>0.7970</td>
<td>0.5585</td>
<td>0.7785</td>
<td>0.6156</td>
<td>0.7265</td>
<td>0.5605</td>
<td>0.6974</td>
<td>0.1151</td>
</tr>
</tbody>
</table>

$n = 100$
manner with perhaps GCV being the next best. They are closer to HQ and HOC in their performance and clearly superior to $R^2$. Interestingly, $R^2$ has the smallest SD for $X1$ and $X2$ but not for $X3$.

For $X2$, MEM has a higher MAPCS than all existing IC. The improvement in MAPCS is clear cut and seems to be achieved by a greater balance in the individual APCSs. On the other hand, MEM has only a slightly higher MAPCS than BIC for $X1$ and $X3$ whether it be the MEM with the largest, modal, or median MAPCS. We also found that for the larger sample sizes, the MEM with the smallest SD has an MAPCS below that of BIC.

MEM has much lower APCSs for the smaller $M_1$ model than BIC for $X2$ and larger APCSs for the larger $M_4$ model. It seems that its clear advantage over BIC comes from being able to adjust the penalties so as not to overfavor $M_1$ with the result that there are larger gains in the APCSs for $M_4$. In the case of $X1$ and $X3$, it appears BIC has reasonably optimal penalties so little can be done by way of changing these penalties to improve the MAPCS.

**5.2 Results when Choosing Between Eight Models**

The relative patterns for $X4$–$X6$ are very similar to those for $X1$–$X3$, particularly with regard to MAPCS values. Again, BIC is the best of the existing IC procedures with $R^2$ being clearly the worst. BIC favors models with fewer parameters while $R^2$ favors models with more parameters. HQ is the second best existing procedure, especially for the larger sample sizes. The largest, median, and modal MAPCS for MEM is always higher than that of BIC, particularly for $X5$. This time we see that MEM and BIC and to a lesser extent HQ are the best procedures. MEM seems to be better placed to exploit any advantage for $n = 20$, but it does appear that as the sample size increases, it is a little harder for MEM to clearly dominate BIC in terms of MAPCS.

**6 CONCLUDING REMARKS**

In this chapter we have outlined a new approach to IC model selection. It is based on the use of simulation to estimate average probabilities of correct selection and choosing penalties that optimize the mean of these average probabilities over the range of models under consideration. It requires the use of the SA algorithm to find these penalties, which can be computationally demanding.

We conducted a Monte Carlo experiment designed to investigate and compare the small sample properties of our new procedure with seven
existing IC procedures. The results of this experiment suggest that often BIC (two out of three cases on average in our limited experience) has close to optimal penalties. Occasionally, and particularly for smaller samples, the new procedure clearly has better small sample properties than BIC. For the best results, care must be taken with the optimization. It may be desirable to try different starting values and not allow boundary values to be restrictive. On the basis of our experiment, we do not recommend the use of $R^2$ for selecting regression models.

Finally, it is worth observing that the approach outlined in this chapter need not be restricted to penalizing maximized likelihood functions. The latter require restrictive distributional assumptions about the models being investigated. It could be that some other estimation criterion such as least squares or least absolute deviations needs to be employed. Clearly, the model selection procedure outlined in this chapter could be readily adapted to such an estimation criterion.

ACKNOWLEDGMENTS

This work was supported by an ARC grant. The authors are grateful to Kwek Kian Teng and Baki Billah for helpful discussions and to Sivagowry Sriananthakumar for research assistance.

REFERENCES


1 INTRODUCTION

The bootstrap, introduced by Efron (1979), is a statistical procedure for estimating the distribution of a given estimator. The distinguishing feature of the bootstrap is that it replaces the unknown population distribution of the data by an estimate of it, which is formed by resampling the original sample randomly with replacement.

When the observed data form a random sample, the bootstrap often provides more accurate critical values for the tests than asymptotic theory (e.g., Singh, 1981; Hall, 1986, 1992; Beran, 1988). Hall (1988) proves that the error in coverage probability made by symmetric two-sided confidence intervals, when bootstrap critical values are used in the IID case, is of order $O(n^{-2})$. This amounts to sizable refinement for the precision of asymptotic confidence intervals, since the errors made by one- and symmetric two-sided asymptotic confidence intervals are of orders $O(n^{-1/2})$ and $O(n^{-1})$. 
respectively.* Monte Carlo experiments support the predictions of the theory, sometimes producing spectacular results (Horowitz, 1994).

In the case of dependent data the bootstrap procedure must be designed in a way that suitably captures the dependence structure of the original sample. Several different sampling procedures have been invented to tackle this task. Carlstein (1986) proposes to divide the original data set in nonoverlapping blocks and then sample these blocks randomly with replacement. Künsch (1989) proceeds similarly, except that he divides the original sample in overlapping blocks. Hall (1985) also suggested these techniques in the context of spatial data. Despite the blocking, the dependence structure of the original sample is not replicated exactly in the bootstrap sample. For example, if nonoverlapping blocks are used, the observations from different blocks in the bootstrap sample are independent with respect to the probability measure induced by bootstrap sampling. Furthermore, observations from the same block are deterministically related. Lastly, the block bootstrap sample is nonstationary even if the original sample is stationary. This dependence structure is unlikely to be present in the original sample. As a result bootstrap performance deteriorates. Hall and Horowitz (1996) give conditions under with Carlstein’s block bootstrap provides asymptotic refinements through $O(n^{-r})$ for coverage probabilities, when bootstrap critical values are used to construct symmetric two-sided confidence intervals for generalized method of moments (GMM) estimators.

The random variables of interest in this chapter are standardized and studentized smooth functions of sample moments of $\hat{X}$ or sample moments of functions of $\hat{X}$, where $\hat{X}$ denotes the sample. For this broad class of random variables we have established the following result: the errors made in the coverage probabilities by one- and symmetric two-sided nonoverlapping block bootstrap confidence intervals are of orders $O(n^{-3/4})$ and $O(n^{-4/3})$, respectively, when optimal block lengths are used. The optimal block lengths are equal to $C_1 n^{1/4}$ and $C_2 n^{1/3}$ for one- and symmetric two-sided confidence intervals, respectively.* Note, however, that the improvement from using the bootstrap over the asymptotics leaves much

---

*By “refinement through $O(n^{-r})$” we mean that the estimated parameter of interest is correct up to and including the term of order $O(n^{-r})$, and the estimation error is of size $o(n^{-r})$.

†For convenience and simplicity we will employ $C$ and/or $C_i$, $i = 1, 2, \ldots$ to denote some finite constants that depend on the specifics of the data-generation process and the functional form of the estimator, but not on sample size, $n$. These constants may assume different values at each appearance. Since the constants do not depend on the sample size, they are not of direct interest given the goals of this paper.
to be desired, especially in the symmetric two-sided confidence interval case. The lackluster performance of the nonoverlapping block bootstrap, as demonstrated in the Monte Carlo experiments of Hall and Horowitz (1996) and Hansen (1999), among others, is consistent with the theory established in the present paper.

To achieve asymptotic refinement, the Edgeworth expansions of the statistic of interest and its bootstrap equivalent have to have the same structure apart from replacing bootstrap cumulants with sample cumulants in the bootstrap expansion.* Lahiri (1992) and Hall and Horowitz (1996) proposed “corrected” bootstrap estimators that achieve asymptotic refinement and partially account for the change in the dependence structure in the bootstrap sample. The corrected versions of bootstrap test statistics are also used in this chapter. The point of the correction factor is to make the exact variance of the leading term of the Taylor series expansion of the bootstrap test statistic equal to one and to do this without introducing new (bootstrap) stochastic terms that would affect the structure of the Edgeworth expansion.

An enlightening fact to note is that one does not need correction factors in the one-sided confidence interval case to achieve asymptotic refinement through $O(n^{-1/2})$ (see, e.g., Lahiri, 1992; Davison and Hall, 1993; Götze and Künsch, 1996; Lahiri, 1996). The reason for this is that the differences between the population and bootstrap variances of higher order terms of the Taylor series expansions of the random variable of interest are of order smaller than $O(n^{-1/2})$.

The solution method used in this chapter was introduced in Hall (1988). A crucial prerequisite for the use of this technique is the existence of the Edgeworth expansions for the random variable of interest. Götze and Hipp (1983, 1994) and Götze and Künsch (1996) give regularity conditions under which the Edgeworth expansions exist for smooth functions of sample averages in the dependent data setting. However, the method of solution of this chapter involves terms that are not smooth functions of sample moments. These terms are of the form $(1/b)^m \sum_{i=1}^{b} \left( \sum_{j=1}^{l} X_{ij} / l \right)^m$, where $m$ is some positive integer, $l$ is the bootstrap block length, $b$ is the number of blocks, and $X_{ij}$ is the $j$th observation in the $i$th block of the bootstrap sample.

---

*An Edgeworth expansion is an approximation to the distribution function of a random variable. Under certain assumptions Edgeworth expansion takes on the form of a power series in $n^{-r}$, where the first term is the standard normal distribution function and $r$ depends on the type of a random variable. The power series form of an Edgeworth expansion makes it a convenient tool for determining the size of the error made by an estimator of a finite sample distribution function. See Hall (1992) for a detailed discussion on Edgeworth expansions.
At present, there are no Edgeworth expansion results in the literature that apply to the statistics of the above type in the dependent data setting. Thus, the techniques used and results obtained in this chapter are heuristic.

The chapter is organized as follows: Section 2 introduces the test statistics of interest, Section 3 lays out the main theoretical results, and Section 4 concludes. This is followed by an Appendix containing the relevant mathematical derivations.

2 TEST STATISTICS

The notation will largely follow that used in Hall et al. (1995) and Hall and Horowitz (1996).

2.1 Sample

The test statistics of interest in this chapter are either normalized or studentized smooth functions of sample moments of $\tilde{X}$ or sample moments of functions of $\tilde{X}$.* Most of the test statistics and estimators in economics are smooth functions of sample averages or can be approximated by such with negligible error. Test statistics based on GMM estimators constitute an example of the latter case (Hall and Horowitz, 1996, Proposition 1, 2).

Denote the data by $\chi = (X_1, \ldots, X_{n_{\text{full}}})$, where $X_i \in \mathbb{R}^d$ is a $d \times 1$ random variable. Assume that $\{X_i\}$ is an ergodic stationary strongly mixing stochastic process and that $E X_i X_j' = 0$ if $|i - j| > k$ for some integer $k < \infty$.† We need to make this assumption so that the estimator of the covariance necessary for studentization would be a smooth function of sample moments. Andrews (1991) proposes a nonparametric covariance matrix estimator, but it is not a smooth function of sample moments and convergence at a rate that is slower than $n^{-1/2}$. Existing theory on Edgeworth expansions with dependent data (Götze and Hipp, 1983, 1994) applies only to smooth functions of sample moments. Set $n = n_{\text{full}} - k$. Define the samples as $\tilde{\chi} \equiv \{\tilde{X}_i; i = 1, \ldots, n\}$, where $\tilde{X}_i = \{X'_i, \ldots, X'_{i+k}\}$. We need to make this redefinition of the sample so that the consistent estimator of the asymptotic variance is a smooth function of sample moments, because the cross-product components of the covariance estimator are not smooth functions of sample moments.

*By “smooth” we mean that the appropriate number of derivatives exist.
†Note that the restriction on the expectations of the products is not equivalent to the assumption of $m$ dependence. Also, this assumption is not as restrictive as it might seem, because $X_i$, for example, can be equal to the GMM moment function. In that case the data can be dependent with an infinite lag, given that some assumptions are satisfied (see Hall and Horowitz, 1996).
2.2 Carlstein’s Blocking Rule

Let \( b \) and \( l \) denote integers such that \( n = bl \). Carlstein’s rule divides the sample \( \tilde{X} \) in \( b \) disjoint blocks, where the \( k \)th block is \( B_k = (\tilde{X}_{(k-1)l+1}, \ldots, \tilde{X}_{kl}) \) for \( 1 \leq k \leq b \). According to the Carlstein rule, bootstrap sample \( \tilde{X}^* \) is formed by choosing \( b \) blocks randomly with replacement out of the set of blocks formed from the original sample and laying the chosen blocks side by side in the order that they are chosen. Bootstrap sample \( \tilde{X}^* \) then consists of \( \{\tilde{X}_i^*\} = \{X_i^*, \ldots, X_{i+l-1}^*\}; i = 1, \ldots, n \).

2.3 Normalized Statistic

Let us denote the random variable of interest by \( U_N = (\hat{\theta} - \theta)/\tilde{s} \), where \( \hat{\theta} = \bar{f}(\tilde{X}) \), \( \theta = \bar{f}(F(X)) \), \( \tilde{s} = [\bar{v}(\hat{\theta} - \theta)]^{1/2} \), \( V(\cdot) \) is an exact variance, and \( \bar{f}(\cdot): \mathcal{R} \rightarrow \mathcal{R} \) is a smooth function of sample moments of \( \tilde{X} \) or sample moments of functions of \( \tilde{X} \).

Let \( U_N^* \) denote the bootstrap equivalent of \( U_N \), where \( U_N^* = (\hat{\theta} - \tilde{\theta})/\tilde{s} \), \( \hat{\theta} = \bar{f}(\tilde{X}^*) \), and \( \tilde{\theta} = \bar{f}(\tilde{X}) \), and \( \tilde{X} = n^{-1}\sum X_i^* \) is the resample mean. Define \( \tilde{s} = (V^\star(\hat{\theta} - \tilde{\theta}))^{1/2} \), where \( V^\star(\hat{\theta} - \tilde{\theta}) = E^\star(\hat{\theta} - \tilde{\theta})^2 \). Here, \( E^\star[\cdot] \) denotes the expectation induced by the bootstrap sampling, conditional on the sample, \( \tilde{X} \). Thus, bootstrap expectation \( E^\star[\cdot] \), and therefore \( \tilde{s} \), can be calculated, since we know the probability distribution induced by bootstrap sampling.

Next, we define the Edgeworth expansions of \( U_N \) and \( U_N^* \):

\[
P(U_N < x) - \Phi(x) - n^{-1/2} p_1(x) - n^{-1} p_2(x) = o(n^{-1})
\]

where \( p_1(x) \) and \( p_2(x) \) are even and odd functions, respectively; both of the functions are polynomials with coefficients depending on cumulants of \( U_N \) and both are of order \( O(1) \).*

\[
P^\star(U_N^* < x) - \Phi(x) - n^{-1/2} \hat{p}_1(x) - n^{-1} \hat{p}_2(x) = o(n^{-1})
\]

except, possibly, if \( \tilde{X} \) is contained in a set of probability \( o(n^{-1}) \). Here \( \hat{p}_1(x) \) and \( \hat{p}_2(x) \) are the same polynomials as above only the population cumulants of \( U_N \) are replaced by sample cumulants of \( U_N^* \), and \( P^\star(\cdot) \) is a

---

*Cumulants are defined as the coefficients of \( (1/\beta)(it)^\gamma \) terms in a power series expansion of \( \log \chi(t) \), where \( \chi(t) \) is the characteristic function of a random variable and \( \chi(t) = \exp(k_1 it + (1/2)k_2(it)^2 + \cdots + (1/\beta)k_{\beta}(it)^\beta + \cdots) \).*
probability measure (conditional on the sample, \(\tilde{\chi}\)) induced by the bootstrap sampling. Let \(k_i\) denote the \(i\)th cumulant of \(U_N\). Then,

\[
\begin{align*}
    n^{-1/2} p_1(x) &= -k_1 - \frac{k_3}{6} (x^2 - 1) \\
    n^{-1} p_2(x) &= -\frac{1}{2} k_1^2 x + \left( \frac{k_4}{24} + \frac{k_1 k_3}{6} \right) (3x - x^3) - \frac{k_3^2}{72} (x^4 - 10x^3 + 15x)
\end{align*}
\]

The first four cumulants of \(U_N\) have the following form (see Appendix, Result 2):

\[
\begin{align*}
    k_1 &= E(U_N) = \frac{k_{1,2}}{n^{1/2}} + \frac{k_{1,3}}{n^{3/2}} + O(n^{-5/2}) \\
    k_2 &= E(U_N - E(U_N))^2 = 1 \\
    k_3 &= E(U_N - E(U_N))^3 = \frac{k_{3,1}}{n^{1/2}} + \frac{k_{3,2}}{n^{3/2}} + O(n^{-5/2}) \\
    k_4 &= E(U_N - E(U_N))^4 \cdot 3[V(U_N)]^2 = \frac{k_{4,1}}{n} + O(n^{-2}),
\end{align*}
\]

where \(k_{i,j}\)'s are constants that do not depend on \(n\) and \(E(U_N^2) = \mathcal{O}(1) + O(n^{-1})\).

Define \(u_a\) as \(P(U_N < u_a) = \alpha\). Inverting the Edgeworth expansion produces the Cornish–Fisher expansion (see Hall, 1992):

\[
\begin{align*}
    u_a - z_a - n^{-1/2} p_{11}(z_a) - n^{-1} p_{21}(z_a) &= o(n^{-1})
\end{align*}
\]

The expansions are to be interpreted as asymptotic series, and in that sense are available uniformly in

\[
\varepsilon < \alpha < 1 - \varepsilon \text{ for any } 0 < \varepsilon < 1/2^*
\]

Similarly, define \(\hat{u}_a\) as \(P^*(U_N^\alpha < \hat{u}_a) = \alpha\). Then, uniformly in \(\varepsilon < \alpha < 1 - \varepsilon\) for each \(\varepsilon > 0\),

\[
\begin{align*}
    \hat{u}_a - z_a - n^{-1/2} \hat{p}_{11}(z_a) - n^{-1} \hat{p}_{21}(z_a) &= o(n^{-1})
\end{align*}
\]

\*In the notation \(p_i(\cdot)\) [and later \(q_i(\cdot)\)], \(i\) denotes the term in the Cornish–Fisher expansion and \(j\) is equal to 1, if \(u_a\) is a percentile of a one-sided distribution, and 2, if it is a percentile of a two-sided distribution.

\*A way to obtain an empirical estimate of \(\hat{u}_a\) is to carry out a Monte Carlo experiment that consists of resampling the original sample \(\tilde{\chi}\), calculating the bootstrap test statistic \(U_N^\alpha\), and forming the empirical distribution of \(U_N^\alpha\) with the desired level of accuracy. The \(\alpha\)th quantile of the empirical distribution of the bootstrap test statistic is the empirical estimate of \(\hat{u}_a\). See Hall (1992) for regularity conditions and details of implementation.
except, possibly, if $\chi$ is contained in a set of probability $o(n^{-1})$. Here,
\[
\begin{align*}
    n^{-1/2}p_{11}(x) &= -n^{-1/2}p_1(x) \\
    n^{-1}p_{21}(x) &= n^{-1/2}p_1(x)n^{-1/2}p'_1(x) - \frac{1}{2}xn^{-1}p_1(x)^2 - n^{-1}p_2(x)
\end{align*}
\]
with obvious modifications for $\hat{p}_{11}(x)$ and $\hat{p}_{21}(x)$. Here, $p'_1(\cdot)$ denotes the derivatives of $p_1(\cdot)$ with respect to $x$.

Let us also introduce some notation for the two-sided distribution function of the normalized test statistic. Noting that $P(|U_N| < x) = P(U_N < x) - P(U_N < -x)$ and that $p_1(x)$ is an even polynomial, the Edgeworth expansions for $|U_N|$ and $|U'_N|$ take on the following form:
\[
\begin{align*}
    P(|U_N| < x) &= 2\Phi(x) + 1 - 2n^{-1}p_2(x) = o(n^{-2}) \\
    P(|U'_N| < x) &= 2\Phi(x) + 1 - 2n^{-1}\hat{p}_2(x) = o(n^{-2})
\end{align*}
\]
where the latter equality holds except, possibly, if $\chi$ is contained in a set of probability $o(n^{-2})$.

Define $\xi = 1/2 \times (1 + \alpha)$, $P(|U_N| < w) = \alpha$, and $n^{-1}p_{12}(\cdot) = -n^{-1}p_2(\cdot)$. Inverting the population Edgeworth expansion we obtain the following Cornish–Fisher expansion:
\[
    w = z_\alpha - n^{-1}p_{12}(z_\alpha) = o(n^{-2})
\]
uniformly in $\varepsilon < \alpha < 1 - \varepsilon$ for each $\varepsilon > 0$. Equivalently, define $P_0(|U'_N| < \hat{w}) = \alpha$ and $n^{-1}\hat{p}_{12}(\cdot) = -n^{-1}\hat{p}_2(\cdot)$, where $\hat{p}_2(\cdot)$ is as $p_2(\cdot)$ with population moments replaced by their sample equivalents. Then, uniformly in $\varepsilon < \alpha < 1 - \varepsilon$ for each $\varepsilon > 0$,
\[
    \hat{w} = z_\alpha - n^{-1}\hat{p}_{12}(z_\alpha) = o(n^{-2})
\]
except, possibly, if $\chi$ is contained in a set of probability $o(n^{-2})$.

### 2.4 Studentized Statistic

The random variable of interest here is $U_S = (\hat{\theta} - \theta)/\hat{s}$, where $\hat{s}^2$ is a consistent estimate of $s^2$. The functional forms of $s^2$ and $\hat{s}^2$ are*
\[
    \hat{s}^2 = \sum_{i=1}^{d} C_i^2 V(\bar{X}_i)
\]
where $V(\bar{X}_i) = \gamma(0)/n + (2/n)\sum_{j=1}^{k}\gamma(j)(1 - n^{-1})$, $\gamma(j)$ is the $j$th auto-covariance of $X$, $k$ is the highest lag for nonzero covariance, $f(\cdot) : \mathcal{R}^d \rightarrow \mathcal{R}$,

*The following variances are expressed in terms of being asymptotically equivalent to something, because, in general, the function $f(\cdot)$ in the random variable of interest, $U_N$, is not a linear function. To be able to evaluate the variance of the random variable of interest, we have to linearize $f(\cdot)$ using Taylor’s theorem.
and $C_i$'s are constants that depend on function $f(\cdot)$, but not on $n$. Also, $\bar{X}_i$ is a sample mean of the $i$th argument of the function $f(\cdot)$. A consistent estimator of $s^2$ is given by
\[ \hat{s}^2 \sim \sum_{i=1}^{d} C_i^2 \hat{V}(\bar{X}_i) \]
where $\hat{V}(\bar{X}_i) = n^{-2} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2 + (2/n) \sum_{m=1}^{k} (1 - n^{-1}m) \sum_{j=1}^{n-m} (X_{ij} - \bar{X}_i) (X_{i,j+m} - \bar{X}_i)/n$ and $X_{ij}$ is the $j$th element of the sample from the $i$th argument.

The corrected bootstrap test statistic is $U_S = (\hat{s}/\bar{s})(\hat{\theta} - \bar{\theta})/\hat{s}$, where $\hat{s}^2$ is the bootstrap equivalent of $s^2$:
\[ \hat{s}^2 \sim \sum_{i=1}^{d} C_i^2 \hat{V}(\bar{X}_i) \]
Here, $\hat{V}(\bar{X}_i) = n^{-2} \sum_{j=1}^{n} (X_{ij}^2 - \bar{X}_i^2)^2 + (2/n) \sum_{m=1}^{k} (1 - n^{-1}m) \sum_{j=1}^{n-m} (X_{ij}^2 - \bar{X}_i^2) (X_{i,j+m}^2 - \bar{X}_i^2)/n$, $X_{ij}$ is the $j$th observation of the $i$th argument of $f(\cdot)$ in the block bootstrap sample, and $\bar{X}_i$ is a sample mean of the block bootstrap sample for the $i$th argument. The exact bootstrap variance of $\hat{\theta} - \bar{\theta}$ is denoted by $\hat{s}^2$:
\[ \hat{s}^2 \sim \sum_{i=1}^{d} C_i^2 \cdot V^*(\bar{X}_i^2 - \bar{X}_i) \]
\[ = \sum_{i=1}^{d} C_i^2 \frac{1}{b} \sum_{j=1}^{b} (\bar{X}_{ij} - \bar{X})^2 \]
\[ = \sum_{i=1}^{d} C_i^2 \frac{1}{n^2} \sum_{j=1}^{b} \sum_{k_1=1}^{l} \sum_{k_2=1}^{l} (X_{ijk_1} - \bar{X})(X_{ijk_2} - \bar{X}) \]
where $V^*(\cdot)$ is the variance induced by block bootstrap sampling, $\bar{X}_{ij}$ is the sample mean of the $j$th block of the $i$th argument, and $X_{ijk_m}$ is the $k_m$th observation in the $j$th block of the $i$th argument.

Note that the Taylor series expansions of $U_S$ and $U_S^*$ have the following forms.*
\[ U_S = U_N \times \left[ 1 - \frac{\hat{s}^2 - s^2}{2s^2} + \frac{3}{8} \left( \frac{\hat{s}^2 - s^2}{s^2} \right) + o_p(n^{-1}) \right] \]
\[ U_S^* = U_N^* \times \left[ 1 - \frac{\hat{s}^2 - \hat{s}^2}{2\hat{s}^2} + \frac{3}{8} \left( \frac{\hat{s}^2 - \hat{s}^2}{\hat{s}^2} \right) + o_p(b^{-1}) \right] \]

*The expansion of $U_S$, of course, is a theoretical construct, since in the studentized case it is assumed that we do not know $s^2$. 
where the error in the second expansion holds conditional on the sample $\chi$. The exact variances of $U_N$ and $U^*_N$ in the above two equations are equal to one. Furthermore, first four cumulants of $U_S$ have the same expansions and rates as the cumulants of $U_N$ above with an exception of the second cumulant. The second cumulant of $U_S$ is equal to $1 + O(n^{-1})$. With this change in the variance, the Edgeworth expansion, say, for $U_S$, is

$$P(U_S < x) = \Phi(x) - n^{-1/2} q_1(x) \phi(x) - n^{-1} q_2(x) \phi(x) = o(n^{-1})$$

where

$$n^{-1/2} q_1(x) = -k^U_1 - k^U_2 \frac{x^2}{2} (x^2 - 1)$$

$$n^{-1} q_2(x) = \left(-k^U_2 - \frac{k^U_1}{2}\right)x + \left(\frac{k^U_4}{24} + \frac{k^U_1 k^U_3}{6}\right)(3x - x^3) - \frac{k^U_5}{72} (x^5 - 10x^3 + 15x)$$

where $k^U_i$ is the $i$th population cumulant of $U_S$, $k^U_2 = 1 + k^U_{2,2}/n + o(n^{-1})$, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and $\phi(\cdot)$ is the standard normal density function. The functional forms of the first two components of one-sided Cornish–Fisher expansion are defined as

$$n^{-1/2} \hat{q}_{11}(x) = -n^{-1/2} q_1(x)$$

$$n^{-1} \hat{q}_{21}(x) = n^{-1/2} q_1(x)n^{-1/2} q'_1(x) - \frac{1}{2} n^{-1} q_1(x)^2 - n^{-1} q_2(x)$$

The functional forms of $n^{-1/2} \hat{q}_{11}(\cdot)$ and $n^{-1} \hat{q}_{21}(\cdot)$ are the same as those of $n^{-1/2} q_{11}(\cdot)$ and $n^{-1} q_{21}(\cdot)$, respectively, with population moments of $U_S$ replaced by the sample cumulants of $U^*_S$. Also, note that the following equality holds for the first polynomial, $n^{-1} q_{12}(\cdot)$, in the Cornish–Fisher expansion of the $a$th quantile of the two-sided population distribution function of the studentized test statistic: $n^{-1} q_{12}(\cdot) = -n^{-1} q_{22}(\cdot)$, with the obvious equivalent for the bootstrap case.

Note the difference between $n^{-1} p_2(\cdot)$ (introduced earlier) and $n^{-1} q_2(\cdot)$. Although the functional forms of the polynomials in the Edgeworth and Cornish–Fisher expansions for the standardized and the studentized statistics are the same (as functions of cumulants), some cancellations happen in the normalized case, when we replace the second cumulant with its expansion. In the normalized case the second cumulant is exactly equal to one, whereas it is equal to $1 + O(n^{-1})$ in the studentized case.
3 MAIN RESULTS

In this section we will establish the optimal nonoverlapping bootstrap block lengths by minimizing the error in the coverage probabilities of one- and symmetric two-sided block bootstrap confidence intervals of normalized and studentized smooth functions of sample averages. Solution methods to the problems involving normalized and studentized statistics are very similar. Section 3.1 deals with the normalized statistic, while the details of the studentized statistic are discussed in Section 3.2. Algebraic details of the important calculations can be found in the Appendix.

3.1 Normalized Statistic

3.1.1 One-sided Confidence Intervals

Here, we find the block length \( \ell \) that satisfies the following expression:

\[
\ell = \arg \min_{\ell \in \mathcal{L}} |P(U_N < \hat{u}_a) - \alpha|
\]

where \( \mathcal{L} \) is the set of block lengths that are no larger than \( n \) and that go to infinity as the sample size \( n \) goes to infinity.

Intuitively, the above probability should equal \( \alpha \) plus some terms that disappear asymptotically and are functions of \( \ell \). The goal, therefore, is to find these approximating terms. We start out by expanding the objective function from the above minimization problem:

\[
P(U_N < \hat{u}_a) = P \left[ U_N - n^{-1/2}(\hat{p}_{11}(z_a) - p_{11}(z_a)) - n^{-1}(\hat{p}_{21}(z_a) - p_{21}(z_a))\right] \\
- \sum_{j=1}^{2} n^{-j/2}p_{j1}(z_a) + z_a + r_N
\]

where \( r_N = o(n^{-1}) \), except, possibly, if \( \chi \) is contained in a set of probability \( o(n^{-1}) \). Let us denote \( n^{-1/2} \Delta_N \equiv n^{-1/2}(\hat{p}_{11}(z_a) - p_{11}(z_a)) \), \( S_N \equiv U_N - n^{-1/2}(\hat{p}_{11}(z_a) - p_{11}(z_a)) - n^{-1}(\hat{p}_{21}(z_a) - \hat{p}_{21}(z_a)) \), and \( p_{j1}(\cdot) \)'s are defined in Eq. (1).

By application of the Delta method (see Appendix, Result 3):

\[
P(U_N < \hat{u}_a) = P(S_N < x) + o(n^{-1})
\]

where \( x = \sum_{j=1}^{2} n^{-j/2}p_{j1}(z_a) + z_a \).

Now the objective is to develop the first four cumulants of \( S_N \) as functions of cumulants of \( U_N \). Then, using cumulants of \( S_N \), to derive an Edgeworth expansion of \( S_N \) as an Edgeworth expansion of \( U_N \) plus some
error terms.* Lastly, evaluate the resulting expression at
\[ x = \sum_{j=1}^{2} n^{-j/2} p_j(z_a) + z_a + z_a. \]

Denote the cumulants of \( S_N \) by \( k_i^S \). Then (see Appendix, Result 2 for details)

\[
\begin{align*}
  k_1^S &= k_1 - n^{-1/2} E(\Delta_N) + o(n^{-1}) \\
  k_2^S &= k_2 - 2n^{-1/2} E(U_N \Delta_N) + o(n^{-1}) \\
  k_3^S &= k_3 - 3n^{-1/2} E(U_N^3 \Delta_N) + 3n^{-1/2} E(U_N^2 \Delta_N) + o(n^{-1}) \\
  k_4^S &= k_4 - 4n^{-1/2} E(U_N^3 \Delta_N) + 12n^{-1/2} E(U_N^2 \Delta_N) + o(n^{-1})
\end{align*}
\]

where we have used the result that \( U_N = O_p(1) \), \( n^{-1/2} \Delta_N = O_p(A_1^{1/2}) \) and \( n^{-1}(\hat{p}_j(z_a) - p_j(z_a)) = O_p(A_1^{1/2}) \) (see Appendix, Result 1), and \( A_1 = C_1 n^{-1/2} + C_2 n^{-1} \Gamma_2 \) and \( A_2 = C_2 n^{-1} \Gamma_2 + C_3 n^{-1} \Gamma_3 \). The rates of \( A_1 \) and \( A_2 \) follow from Hall et al. (1995). Next, substitute these cumulants in the Edgeworth expansion of \( S_N \). The resulting equation is

\[
P(U_N \leq \bar{u}_a) = P(S_N \leq x) + o(n^{-1})
\]

\[
= P(U_N \leq x) + n^{-1/2} E(\Delta_N)\phi(x) + n^{-1/2} E(U_N \Delta_N)\phi(x) \\
+ \left( \frac{1}{2} n^{-1/2} E(U_N^2 \Delta_N) - \frac{1}{2} n^{-1/2} E(U_N^3 \Delta_N) \right) (x^2 - 1) \phi(x) \\
+ \left( \frac{1}{2} n^{-1/2} E(U_N^2 \Delta_N) - \frac{1}{2} n^{-1/2} E(U_N^3 \Delta_N) \right) \\
\times (3x - x^3) \phi(x) + o(n^{-1})
\]

Evaluating the above equation at \( x = \sum_{j=1}^{2} n^{-j/2} p_j(z_a) + z_a \) and noting that \( P(U_N \leq x) = \alpha + O(n^{-1}) \) does not depend on the block length, \( l \), gives us the following objective function:

\[
n^{-1/2} \sum_{i=0}^{3} E(U_N^i \Delta_N)C_i + n^{-1/2} E(U_N^3 \Delta_N)C_4 \\
+ n^{-1/2} E(U_N^3 \Delta_N)C_5 + o(n^{-1})
\]

where \( n^{-1/2} E(U_N^i \Delta_N) - n^{-1/2} E(U_N^i \Delta_N) \), \( \{i = 0, 1\} \). Thus, we are left with four terms: \( n^{-1/2} E(U_N^i \Delta_N) \), \( \{i = 0, \ldots, 3\} \). Result 4 in the Appendix

---

*In this chapter we have not derived the regularity conditions under which this expansion exists.
shows that these terms have the following orders:

\[
\begin{align*}
    n^{-1/2} E(\Delta_N) &= O(n^{-3/2} l^{1/2}) + O(n^{-1} l^{-1}) \\
    n^{-1/2} E(U_N \Delta_N) &= O(n^{-1} l) + O(n^{-1} l^{-1}) \\
    n^{-1/2} E(U^2_N \Delta_N) &= O(n^{-3/2} l^{1/2}) + O(n^{-1} l^{-1}) \\
    n^{-1/2} E(U^3_N \Delta_N) &= O(n^{-1} l^{3/2}) + O(n^{-1} l^{-1})
\end{align*}
\]

Therefore, the error in the bootstrap coverage probability of a one-sided block bootstrap confidence interval is: \( O(n^{-1} l) + O(n^{-1} l^{-1}) \). The block length, \( l \), that minimizes this quantity is proportional to \( n^{1/4} \). Furthermore, the size of the coverage error is \( (n^{-3/4}) \), when block lengths proportional to \( n^{1/4} \) are used.

### 3.1.2 Symmetric Two-Sided Confidence Intervals

The solution methods for one- and symmetric two-sided confidence interval cases are very similar. Again, we are looking for the block length, \( l \), that satisfies the following equation:

\[
l^* = \min_{l \in \mathcal{L}} |P(|U_N| < \hat{w}_a) - \alpha|
\]

Note that \( \hat{w}_a = w_a = n^{-1} \Delta_N^A + o(n^{-2}) \), except, if \( \chi \) is contained in a set of probability \( o(n^{-2}) \), where \( n^{-1} \Delta_N^A = n^{-1} (\bar{p}_{12}(z_1) - p_{12}(z_1)) \) and \( n^{-1} p_{12}(\cdot) = n^{-1} \tilde{p}_{12}(\cdot) \). One can show (see Appendix, Result 1) that \( n^{-1} \Delta_N^A = O_p(A_2^{1/2}) \), where \( A_2 = C_1 n^{-2} l^{-2} + C_1 n^{-2} l^{-2} + C_2 n^{-3} l^2 \) and the rate of \( A_2 \) follows from Hall et al. (1995).

Then:

\[
\begin{align*}
    P(|U_N| < \hat{w}_a) &= P(|U_N| < w_a + n^{-1} \Delta_N^A + r_N^A) \\
    &= P(|U_N| < w_a + n^{-1} \Delta_N^A) + O(n^{-2}) \\
    &= P(U_N < w_a + n^{-1} \Delta_N^A) - P(U_N < w_a + n^{-1} \Delta_N^A) \\
    &\quad + O(n^{-2})
\end{align*}
\]

where \( r_N^A = o(n^{-2}) \), except, possibly, if \( \chi \) is contained in a set of probability \( o(n^{-2}) \) and the second equality follows by the Delta method (see Appendix, Result 3). The next task is to develop cumulants of \( U_N - n^{-1} \Delta_N^A \) and \( U_N + n^{-1} \Delta_N^A \) and substitute them in the Edgeworth expansion of

\[
P(U_N - n^{-1} \Delta_N^A < w_a) - P(U_N + n^{-1} \Delta_N^A < -w_a).
\]

Following the steps of the solution method for the one-sided confidence interval case, it is straightforward to show that the relevant error terms are: \( n^{-1} E(\Delta_N^A) \),
n^{-1}E(U_N^2 \Delta_N^4), and n^{-1}E(U_N^2 \Delta_N^4)E(U_N)\tilde{n}^{-1}E(U_N \Delta_N^4). The above terms are of the following orders (see Appendix, Result 4 for the methods used):

\[ n^{-1}E(\Delta_N^4) = \mathcal{O}(n^{-2}l^2) + \mathcal{O}(n^{-1}l) \]
\[ n^{-3/2}E(U_N \Delta_N^4) = \mathcal{O}(n^{-2}l^{3/2}) + \mathcal{O}(n^{-2}l^{-1}) \]
\[ n^{-1}E(U_N^2 \Delta_N^4) = \mathcal{O}(n^{-2}l^2) + \mathcal{O}(n^{-1}l^{-1}) \]

Thus, the error in the coverage probability of a symmetric two-sided block bootstrap confidence interval is of order \( \mathcal{O}(n^{-2}l^2) + \mathcal{O}(n^{-1}l^{-1}) \). The block length, \( l \), that minimizes this error is proportional to \( n^{1/3} \). The error then is of size \( \mathcal{O}(n^{-4/3}) \).

### 3.2 Studentized Statistic

It is intuitively clear that the error rates of the coverage probability in the studentized case should be the same in the normalized case. The reason for this is that the Taylor series expansion of the studentized test statistic is equal to normalized test statistic plus some higher order error terms [see Eq. (2)].

The solution method for the studentized statistic case is very similar to that of the normalized statistic case for both one- and symmetric two-sided confidence intervals. The dominant error terms are: \( n^{-1}E(U_S^2 \Delta_S^4), \{i = 0, \ldots, 3\} \) for the one-sided case and \( n^{-1}E(\Delta_S^4), n^{-1}E(U_S^2 \Delta_S^4), \) and \( n^{-1}E(U_S^2 \Delta_S^4)E(U_S) \) for the two-sided case, where \( n^{-1/2} \Delta_S = n^{-1/2}(q_{11}(z) - q_{12}(z)) \) and \( n^{-1} \Delta_S^4 = n^{-1}(q_{12}(z_2) - q_{12}(z_2)) \) [see Equation (4)].

Let \( k_i^U \) and \( \hat{k}_i^U \) denote the population and bootstrap cumulants of \( U_S \) and \( U_N \), respectively. Given the structure of the polynomials \( q_1(\cdot) \) and \( q_2(\cdot) \) in Eq. (3) we see that the following errors terms have to be bounded: for the one-sided case, \( E[U_S^2(k_i^U - k_i^U)] \) and \( E[U_S^2(k_i^U - k_i^U)] \), \{j = 0, \ldots, 3\}; for the symmetric two-sided case: \( E[U_S^2(k_i^U - k_i^U)], E[U_S^2(k_i^U - k_i^U)], E[U_S^2(k_i^U - k_i^U)], E[U_S^2(k_i^U - k_i^U)], \{j = 0, \ldots, 2\} \).

Notice that the above terms are dominated by their normalized statistic equivalents. This is easy to see from Eq. (2), where we break down \( U_S \) and \( U_N \) in \( U_N \) and \( U_N^* \), respectively, times something that is asymptotically equal to one. The only exception occurs in the case of the terms \( E[U_S^2(k_i^U - k_i^U)], \{j = 0, \ldots, 2\} \). In the normalized statistic case the exact variances of \( U_N \) and \( U_N^* \) are both equal to one. Thus, the leading terms of \( k_i^U \) and \( \hat{k}_i^U \) both cancel.

*Note that all the error terms above are nonstochastic. Thus, the word “dominance” is used in its usual limit sense. When the terms involved are stochastic, a specific reference to stochastic dominance is made.
and $E[U_j(\hat{k}_2^U - k_2^U)]$ is dominated by the population and the bootstrap variances of the second brackets in Eq. (2). However, one can show (see Appendix) that $E[U_j(\hat{k}_2^U - k_2^U)]$ terms are either equal to or dominated by $O(n^{-1/4})$. It follows that the error rates in the coverage probabilities of one- and symmetric two-sided block bootstrap confidence intervals of the studentized statistics are $O(n^{-1}) + O(n^{-1/2}I^{-1})$ and $O(n^{-2}I^{-1}) + O(n^{-1}I^{-1})$, respectively. Thus, the optimal block lengths and the coverage error rates are the same for both studentized and normalized cases, where the former are proportional to $n^{1/4}$ and $n^{1/3}$ for one- and symmetric two-sided confidence intervals, respectively.

4 CONCLUSIONS

In this chapter we have established that the minimum error rates in coverage probabilities of one- and symmetric two-sided nonoverlapping block bootstrap confidence intervals are of orders $O(n^{-3/4})$ and $O(n^{-4/3})$, respectively, for normalized and studentized smooth functions of sample moments. These rates are attained when the blocks for one- and symmetric two-sided block bootstrap confidence intervals are proportional to $n^{1/4}$ and $n^{1/3}$, respectively. The above rates in the coverage errors are consistent with the Monte Carlo evidence of Hall and Horowitz (1996) and Hansen (1999). The reason for such slight refinement over the asymptotic case is clear: the blocking damages the dependence structure of the original sample, and even under the optimal blocking rate, the block bootstrap does not recover this dependence sufficiently well.

Bühlman (1997, 1998) has suggested a promising alternative to blocking called “sieve bootstrap,” which seems to work well under certain restrictions.

ACKNOWLEDGMENT

I thank Professor Joel Horowitz for his indispensable comments and continued support. All of the errors are my own.

REFERENCES


Carlsen, E., 1986. The use of subseries methods for estimating the variance of a
1179.
implementing the bootstrap with dependent data. *Australian Journal of
random vectors. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte
Gebiete*, 64, 211–239.
20, 231–246.
1431–1452.
Statistical Society*, 50, 35–45.
Hall, P. and Horowitz, J.L., 1996. Bootstrap critical values for tests based on
with dependent data. *Biometrika*, 82, 561–574.
Hansen, B., 1999. Non-parametric dependent data bootstrap for conditional
moment models. Working paper, Dept. of Economics, University of
Wisconsin, Madison, Wisconsin.
Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary
and nonstationary data. In: *Exploring the Limits of Bootstrap*, R. Lepage and
Lahiri, S.N., 1996. On Edgeworth expansion and moving block bootstrap for
studentized M-estimators in multiple linear regression models. *Journal of
Politis, D. and Romano, J., 1994. Large sample confidence regions based on

**APPENDIX**

**Result 1** Derivation of the probability bounds for $n^{-1/2} \Delta_N$, $n^{-1} \Delta_N^A$ and $n^{-1} (\hat{p}_{21}(x) - p_{21}(x))$.

From Hall et al. (1995), we know that $n^{-1} E(\hat{p}_1(x) - p_1(x))^2 = O(A_1)$, where $A_1 = C_1 n^{-1} l^2 + C_2 n^{-2} \bar{p}$. Also, note that the probability rate of $n^{-1/2} (\hat{p}_1(x) - p_1(x))$ is same as that of $n^{-1/2} (\hat{p}_{11}(x) - p_{11}(x))$, since $n^{-1/2} p_{11}(.) = -n^{-1/2} p_1(.)$ with the obvious modifications for $n^{-1/2} \hat{p}_{11}()$. Then, by Chebyshev’s inequality:

$$P(A_1^{-1/2} n^{-1/2} | \Delta_N| > M_\epsilon) < \frac{n^{-1} E(\Delta_N)^2}{A_1 M_\epsilon^2} \equiv \epsilon^*$$

where $M_\epsilon < \infty$ and $\epsilon^*$ can be made arbitrarily small. The latter statement is true, because $n^{-1} E(\Delta_N)^2 / A_1 = O(1)$. Thus, $A_1^{-1/2} n^{-1/2} \Delta_N = O_p(1)$, i.e., it is bounded in probability.

To find the probability bound for $n^{-1} \Delta_N^A$ we use the result from Hall et al. (1995): $n^{-2} E(\hat{p}_2(x) - p_2(x))^2 = O(A_2)$, where $A_2 = C_1 n^{-2} l^2 + C_2 n^{-3} \bar{p}^3$. Also, the probability rate of $n^{-1} (\hat{p}_2(x) - p_2(x))$ is the same as that of $n^{-1} (\hat{p}_{12}(x) - p_{12}(x))$, since $n^{-1} p_{12}(.) = -n^{-1} p_2(.)$ with the obvious modifications for $n^{-1} \hat{p}_{12}()$. We then follow the steps above to establish that $n^{-1} \Delta_N^A = O_p(A_2^{-1/2})$.

Lastly, to establish the probability bound of $n^{-1} (\hat{p}_{21}(x) - p_{21}(x))$, we note that the probability rate of $n^{-1} (\hat{p}_2(x) - p_2(x))$ is the same as that of $n^{-1} (\hat{p}_{22}(x) - p_{22}(x))$ (this is not hard to show), and then proceed as in the case above.

**Result 2** Derivation of the cumulants of $U_N$, $S_N$, and $U_N \pm n^{-1} \Delta_N^A$

The derivation of the cumulants of $U_N$ depend on applying the Taylor series expansion to the random variable of interest. We know that

$$U_N = \frac{n^{1/2} (f(X) - f(\mu))}{\sqrt{V[n^{1/2} (f(X) - f(\mu))]}$$
Note that $V[n^{1/2}(f(\hat{X}) - f(\mu))] = \mathcal{O}(1)$. Then, using the Taylor expansion with respect to $\hat{X}$ around $\mu$:

$$n^{1/2}( f(\hat{X}) - f(\mu)) = \sum_{i=1}^{d} (D_i f)(\mu)n^{1/2}(\hat{X}_i - \mu_i) + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} (D_i D_j f)(\mu)n^{1/2}(\hat{X}_i - \mu_i)(\hat{X}_j - \mu_j) + o(n^{-1/2})$$

where the notation is as in Eq. (2). Then

$$E(U_N) = k_{1,2} \frac{n^{1/2}}{\sqrt{n}} + k_{1,3} \frac{n^{3/2}}{\sqrt{n}} + \mathcal{O}(n^{-5/2})$$

where $k_{i,j}$ are constants that do not depend on $n$. Here, we have used the following equalities from Hall et al. (1995):

$$E(\hat{X}_i - \mu_i)^2 = \frac{\gamma(0)}{n} + 2 \frac{n}{\sqrt{n}} \sum_{j=1}^{k} (1 - n^{-1}) \gamma(j)$$

(5)

$$E(\hat{X}_i - \mu_i)^3 = \frac{E(X_i - \mu_i)^3}{n^2} + \frac{3}{n} \sum_{j=1}^{k} (1 - n^{-1}) j \gamma(j) E((X_0 - \mu_i)(X_{ij} - \mu_i)^2$$

$$+ (X_0 - \mu_i)^2 (X_{ij} - \mu_i)) + \frac{6}{n^2} \sum_{j_2,j_3 \geq 1} \sum_{j_2+j_3 \leq k} (1 - n^{-1}) (j_2 + j_3)$$

$$\times E((X_0 - \mu_i)(X_{ij_2} - \mu_i)(X_{ij_2+j_3} - \mu_i))$$

(6)

where $X_{ij}$ is the $j$th observation of the $i$th element of the vector $X$, $\gamma(j)$ is the lag-$j$ covariance of the $i$th element of the vector $X$. Also, we used the moment inequalities of Yokoyama (1980), Doukhan (1994) (Remark 2, p. 30), Hölder, and a consequence of Hölder and Burkholder inequalities (Hall and Heyde, 1980, eq. 3.67, p. 87). These tools will be used repeatedly when we bound the error terms in the coverage probability (see Results 4 and 5).

To derive higher order cumulants, use the Taylor series expansion, taken to the appropriate power.

The method of derivation of cumulants of $S_N$ and $U_N \pm n^{-1} \Delta_N A$ is to derive them as sums of cumulants of $U_N$ plus an error that is asymptotically equal to zero. Let us demonstrate this for the second cumulant of $S_N$:

$$k_2^S = E(S_N)^2 - E^2(S_N)$$

$$= E(U_N - n^{-1/2} \Delta_N - n^{-1}(\hat{p}_{21}(z_0) - p_{21}(z_0)))^2$$

$$- E^2(U_N - n^{-1/2} \Delta_N - n^{-1}(\hat{p}_{21}(z_0) - p_{21}(z_0)))$$

$$= k_2 - 2n^{-1/2} E(U_N \Delta_N) + n^{-1} E(\Delta_N^2)$$

$$+ 2n^{-1/2} E(U_N) E(\Delta_N) - n^{-1} E^2(\Delta_N) + o(n^{-1})$$
Using this method it is straightforward to derive cumulants of higher orders.

**Result 3** Derivations involving the Delta method

Here, we will demonstrate the derivation of equality $P(U_N < \hat{u}_a) = P(S_N < x) + o(n^{-1})$. The derivation of other equalities involving applications of Delta method are similar.

$$P(U_N < \hat{u}_a) = P(S_N < x + r_N)$$

where $r_N = o(n^{-1})$, except, possibly, if $Y$ is contained in a set of probability $o(n^{-1})$. That is, $P(r_N = o(n^{-1})) = o(n^{-1})$. Therefore, as $n \to \infty$, $P(m|r_N) \geq \varepsilon$, for some $m \geq n = o(n^{-1})$, for all $\varepsilon > 0$. This is equivalent to $P(n|r_N) \geq \varepsilon) = o(n^{-1})$ for all $\varepsilon > 0$, as $n \to \infty$. Then,

$$P(S_N < x + r_N) = P \left( S_N \leq x + r_N, \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{ \omega : m \cdot |r_N(\omega)| < \varepsilon \} \right)$$

$$+ P \left( S_N < x + r_N, \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{ \omega : m \cdot |r_N(\omega)| \geq \varepsilon \} \right)$$

$$\leq P \left( S_N < x + r_N, \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{ \omega : m \cdot |r_N(\omega)| < \varepsilon \} \right)$$

$$+ P \left( \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{ \omega : m \cdot |r_N(\omega)| \geq \varepsilon \} \right)$$

$$\leq \lim_{n \to \infty} P(S_N < x + r_N, m \cdot |r_N| \leq \varepsilon, \forall m \geq n)$$

$$+ P((\omega : r_N(\omega) \neq o(n^{-1})))$$

$$\leq P(S_N < x) + o(n^{-1})$$

Similarly, we can show that $P(S_N < x) + r_N) \geq P(S_N < x) + o(n^{-1})$.

**Result 4** Bounding of $n^{-1/2} E(U_N^i, \Delta_N), \{ i = 0, \ldots, 3 \}$

$$n^{-1/2} E(\Delta_N^i) = n^{-1/2} E(\hat{\Delta}_i - \Delta_i)$$

$$= E(\hat{\Delta}_1 - k_1) + C \cdot E(\hat{\Delta}_3 - k_3).$$
Start with $E(\hat{k}_1 - k_1)$ and define $\hat{\beta} \equiv E^*(f(\bar{X}^*) - f(\bar{X}))$, $\beta \equiv E(f(\bar{X}) - f(\mu))$, and $\hat{k}_1 = \hat{\beta}/\hat{s}$. Then,

$$
k_1 = \frac{\hat{\beta} - \beta + \beta}{s} \left( 1 - \frac{s^2}{2s^2} + \frac{3}{8} \left( \frac{s^2}{s^2} \right)^2 + \cdots \right)$$

$$= k_1 + \frac{\hat{\beta} - \beta}{s} \frac{s^2}{s^2} - s^2 - \frac{\hat{\beta} - \beta}{s^2} \frac{s^2}{s^2} - \frac{\beta}{s} \frac{s^2}{s^2} - \frac{3}{8} \left( \frac{s^2}{s^2} \right)^2 \mathcal{O} (n^{3/2} A_0)
$$

where $A_0 = C_1 n^{-2} l^{-2} + C_2 n^{-3} l$, $s^2 = \mathcal{O}(n^{-1})$, $\beta = \mathcal{O}(n^{-1})$, $\hat{\beta} - \beta = \mathcal{O}(A_0^{1/2})$, \hat{s} = \mathcal{O}(A_0^{1/2})$, $E(\hat{\beta} - \beta) \sim C_1 n^{-1} l^{-1} + C_2 n^{-2} l \sim E(\hat{s}^2 - s^2)$, $E(\hat{s}^2) \sim C_1 n^{-2} l^{-2} + C_2 n^{-3} l \sim E(\hat{s}^2 - s^2)^2$.

The last six bounds are from Hall et al. (1995). Also, note that $\hat{\beta} - \beta \sim (f'(\bar{X})/2) E(X_i^2 - X_i)^3 - (f'(\mu)/2) E(X_i - \mu)^3$ and $\hat{s}^2 - s^2 \sim (f'(\bar{X}))^2 E(X_i^3 - X_i^3) - (f'(\mu))^2 E(X_i - \mu)^3$. Then, $E(\hat{k}_1 - k_1) \sim C_1 n^{-1/2} l^{-1} + C_2 n^{-3/2} l$.

Next, bound $E(\hat{k}_3 - k_3)$:

$$E(\hat{k}_3 - k_3) = E\left( \frac{f(\bar{X}) - f(\bar{X})}{\sqrt{V(f(\bar{X}))}} - E\left( \frac{f(\bar{X}) - f(\bar{X})}{\sqrt{V(f(\bar{X}))}} \right) \right)^3$$

$$- E\left( \frac{f(\bar{X}) - f(\mu)}{\sqrt{V(\bar{X})}} - E\left( \frac{f(\bar{X}) - f(\mu)}{\sqrt{V(\bar{X})}} \right) \right)^3$$

$$\sim C \left( \frac{E(\bar{X}^3 - \bar{X}_i^3)}{\sqrt{V(\bar{X}_i - \bar{X}_i)}} \right)^3 \left( \frac{E(\bar{X}_i - \mu)^3}{(V(X_i - \mu)_{3/2})^2} \right)$$

Note that $\bar{X}$ is a vector random variable and $\bar{X}_i$ is a scalar random variable. Here, we have used Taylor's theorem for vector-valued functions. By Hall et al. (1995):

$$E\left( \frac{E(\bar{X}^3 - \bar{X}_i^3)}{\sqrt{V(\bar{X}_i - \bar{X}_i)}}^3 \right) \left( \frac{E(\bar{X}_i - \mu)^3}{(V(X_i - \mu)_{3/2})^2} \right)$$

$$\sim \mathcal{O}(n^{-1/2} l^{-1}) + \mathcal{O}(n^{-3/2} l^{1/2})$$

Therefore, $n^{-1/2} E(\Delta_N) \sim C_1 n^{-3/2} l^{1/2} + C_2 n^{-1/2} l$. 

$$n^{-1/2} E(U_N \Delta_N) = E(U_N \hat{k}_1) - k_1 + C \cdot (E(U_N \hat{k}_3) - k_1 k_3)$$
Since $\hat{k}_1 = k_1 + (\hat{\beta} - \beta)/s - (\beta/s)((\hat{s}^2 - s^2)/2s^2) + O_p(n^{3/2}A_0)$:

$$E(U_X\hat{k}_1) = k_1^2 + \frac{1}{s^2}E((f(\hat{X}) - f(\mu))(\hat{\beta} - \beta))$$

$$+ E \left( \frac{\hat{U}_N}{N} \frac{\beta \hat{s}^2 - s^2}{2s^2} \right) + o(A_3)$$

where $A_3 = C_4n^{-1} + C_5n^{-1/2}$. The rate of the error $o(A_3)$ stems from the following two considerations. First, the terms covered by the error are farther out in the Taylor series expansion of $\hat{k}_1$ than the terms left in the expansion, and therefore their rates are smaller than those of the terms left in the expansion. Second, the error of the term $n^{-1/2}E(U_X\Delta_X)$ turns out to be $O(A_3)$. Next, define $Y(X - \mu$, i.e., $Y$ is the demeaned random vector $X$.

Some algebra:

$$(f(\hat{X}) - f(\mu))(\hat{\beta} - \beta) \sim f'(\mu)(\hat{X}_i - \mu_i)$$

$$\times \left[ \frac{f''(\mu)}{2b} \sum_{j=1}^{b}(X_{ij} - \hat{X}_i)^2 - \frac{f''(\mu)}{2} E(\hat{X}_i - \mu_i)^2 \right]$$

$$= C \left( \tilde{Y}_b^{-1}(\tilde{Y}_i^{(2)} - \tilde{Y}_i) - \tilde{Y}_i \frac{1}{b} E(\tilde{Y}_i^{(2)}) \right)$$

$$+ C \left( \tilde{Y}_b^{-1} E(\tilde{Y}_i^{(2)}) - \tilde{Y}_i E(\tilde{Y}_i^{(2)}) \right)$$

where $X_{ij} = (1/b)\sum_{k=1}^{b}X_{i,(j-1)k+k}$, $\tilde{Y}_i^{(k)} = (1/b)\sum_{j=1}^{b}(X_{ij} - \mu_i)^k$, and $C = f'(\mu) \times f(\mu)/2$.

$$E[(f(\hat{X}) - f(\mu))(\hat{\beta} - \beta) \sim \frac{C}{b} E[\tilde{Y}_b^{-1}(\tilde{Y}_i^{(2)} - \tilde{Y}_i) - \tilde{Y}_b E(\tilde{Y}_i^{(2)})]$$

$$= \frac{C}{b} E[\tilde{Y}_b^{-1}(\tilde{Y}_i^{(2)} - E(\tilde{Y}_i^{(2)})) - \tilde{Y}_i^3]$$

$$= O(n^{-2}) + O(n^{-3}l)$$

where the second equality follows from Taylor’s theorem and the last equality follows from the application in equalities of Yokoyama (1980), Doukhan (1994) (Remark 2, p. 30), and Eq. (6). By noting that $E[U_X(\hat{\beta} - \beta)/s] \sim E[U_X((\hat{s}^2 - s^2)/2s^2)]$, it follows then that $E(U_X\hat{k}_1) - k_1^2 = O(n^{-1}) + O(n^{-3}l)$. 
Let us examine $E(U_N \hat{k}_3) - k_1k_3$. From Hall et al. (1995), $\hat{k}_3 = k_3 + (l^{1/2}/n^{1/2})k_3 - k_3 + O_p(A_l^{1/2})$, where $A_l = C_1n^{-1}l^{-2} + C_2n^{-1}l^2$ and $k_3$ is the third cumulant for a sample with $l$ observations. Then,

$$E(U_N \hat{k}_3) - k_1k_3 \sim k_1 \left( \frac{l^{1/2}}{n^{1/2}}k_3 - k_3 \right) + E(U_N R_N)$$

$$= \frac{k_1}{n^{1/2}} \cdot O(l^{-1}) + O(n^{-1}l)$$

$$= O(n^{-1}l^{-1}) + O(n^{-1}l)$$

where $R_N = O(n^{-1/2}l^3)(\bar{Y}_i^{(3)} - E(\bar{Y}_i^{(3)})) + O(n^{-1/2}l)\bar{Y}_i + O(n^{-1/2}l)(\bar{Y}_i^{(2)} - E(\bar{Y}_i^{(2)}))$. The first part of the second line of the above equation follows from Hall et al. (1995), and the second part of the second line follows from inequalities of Yokoyama (1980), Doukhan (1994) (Remark 2, p. 30), and Eqs. (5) and (6).

Thus, $n^{-1/2}E(U_N\Delta_N) = O(n^{-1}l^{-1}) + O(n^{-1}l^{-1})$. Following methods developed above we can establish that $n^{-1/2}E(U_N^2\Delta_N) = O(n^{-3/2}l^{3/2}) + O(n^{-1/2}l^{-1})$ and $n^{-1/2}E(U_N^3\Delta_N) = O(n^{-1}l^{1/2}) + O(n^{-1}l^{-1})$. Therefore, the error in coverage probability of the one-sided confidence interval is equal to $O(n^{-1}l^{-1}) + O(n^{-1}l)$.

To obtain the rate of the error in coverage probability for the symmetric two-sided confidence interval, we have to bound the following terms: $n^{-1}E(\Delta_N^2)$, $n^{-3/2}E(U_N\Delta_N)$, and $n^{-1}E(U_N^2\Delta_N^2)$. Applying the methods above we can show that $n^{-1}E(\Delta_N^2) = O(n^{-2}l^2) + O(n^{-1}l^{-1})$, $n^{-3/2}E(U_N\Delta_N) = O(n^{-2}l^2) + O(n^{-1}l^{-1})$, and $n^{-1}E(U_N^2\Delta_N^2) = O(n^{-2}l^2) + O(n^{-1}l^{-1})$.

**Result 5** Bounding of $E[U_j^*(V_N^*(U_S^3) - V(U_S))]$, $\{j = 0, \ldots, 2\}$

$$E[V^*(U_S^3) - V(U_S)] \sim CE \left\{ E^* \left( \frac{f'(\bar{X}_i)\bar{X}_i^2}{s} \right) \right.$$  

$$- E \left( \frac{f'(\mu_i)\bar{X}_i - \mu_i}{s} \right) \left( \frac{s^2 - \bar{s}^2}{2s^2} \right) \right\}$$

$$\sim O(n)E[D^2(\bar{X}_i^2 - \bar{X}_i^2)(\bar{X}_i^{(2)} - \bar{X}_i^{(2)})]$$

$$- E((\bar{X}_i - \mu_i)^2(\bar{X}_i^{(2)} - E(X_i^{(2)}))]$$

$$+ O(n)E[D^2((\bar{X}_i^2 - \bar{X}_i^2)(\bar{X}_i^{(2)} - \bar{X}_i^{(2)})]$$

$$- E((\bar{X}_i - \mu_i)^2(\bar{X}_i^{(2)} - \bar{X}_i^{(2)}))]$$
\begin{align*}
\sim & \mathcal{O}(n) E\left( (\bar{X} - X)^3 E(X_i - \mu_i)^3 \right) \\
& + \mathcal{O}(n) E\left\{ \frac{1}{b^3} E(\bar{X}_{ij} - \bar{X}_i)^4 - E(\bar{X}_i - \mu_i)^4 \right\} \\
\sim & C E \left\{ \frac{\hat{p}^3}{n} \sum_{j=1}^b \frac{(\bar{X}_{ij} - \mu_i)^3}{b} - n E(\bar{X}_i - \mu_i)^3 \right\} \\
= & C \left( \frac{\hat{p}^3}{n} E(\bar{X}_{ij} - \mu_i)^3 - n E(\bar{X}_i - \mu_i)^3 \right)
\end{align*}

where $E^*(\bar{X} - X)^3 = (1/b^2) \sum_{i=1}^b (\bar{X}_i - X)^3 / b$, $s^2 = \mathcal{O}(n^{-1})$, and the rest of the notation is as in Result 4. Then, by Eq. (6), we have that $\hat{p}^3/n E(X_{ij} - \mu_i)^3 - n E(\bar{X}_i - \mu_i)^3 = \mathcal{O}(n^{-1} I^{-1})$. Using the above methodology it is straightforward to show that the terms $E[U_3(k_{ij}^U - k_2^U)]$ and $E[U_2(k_{ij}^U - k_2^U)]$ are dominated by the term $E[k_{ij}^U - k_2]$. 

A Comparison of Alternative Causality and Predictive Accuracy Tests in the Presence of Integrated and Cointegrated Economic Variables

Norman R. Swanson
Rutgers University, New Brunswick, New Jersey, U.S.A.

Ataman Ozyildirim
The Conference Board, New York, New York, U.S.A.

Maria Pisu
University of Alabama, Birmingham, Alabama, U.S.A.

1 INTRODUCTION

In a 1983 paper, Geweke et al. compared eight alternative tests of the absence of causal ordering, in the context of stationary variables. All were versions of the tests proposed by Granger (1969) and Sims (1972). Their unambiguous finding was that Wald variants of the tests are preferred. This chapter is meant to extend the results of Geweke et al. (1983) in two directions. First, cases where the variables are not I(0) [in the sense of Engle and Granger (1987)], but also integrated of higher order and/or cointegrated, are considered. Second, in addition to classical in-sample tests of Granger noncausality, various model selection procedures for checking causal ordering are outlined and examined.
Our results, based on a series of Monte Carlo experiments, indicate that surplus lag Wald-type tests due to Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996) (hereafter the TYDL test) perform well, regardless of integratedness and/or cointegratedness properties. However, when cointegrated data are examined, the sequential Wald-type tests due to Toda and Phillips (1993, 1994) come in a close second, and in some cases are superior. In addition, we find that model selection procedures based on the use of Schwarz information criteria as well as on the \textit{ex ante} forecasting approach suggested in Diebold and Mariano (1995) and Swanson (1998) perform surprisingly well in a variety of different contexts.

It should perhaps be noted that out-of-sample predictive ability type tests such as the model selection-based \textit{ex ante} procedure that we consider here are receiving considerable attention in the current causality literature. One reason for this may be an aspect of Granger’s original definition which had not previously received much attention in the literature. In particular, at issue is whether or not standard in-sample implementations of Granger’s definition are wholly in the spirit originally intended by Granger, and whether out-of-sample implementations might also be useful. Arguments in favor of using out-of-sample implementations are given in Granger (1980, 1988), and are summarized nicely in Ashley et al. (1980), where it is stated on page 1149 that: “. . . a sound and natural approach to such tests [Granger causality tests] must rely primarily on the out-of-sample forecasting performance of models relating the original (non-prewhitened) series of interest.” Recent papers that develop and discuss predictive ability-type causality tests include Clark and McCracken (1999), McCracken (1999), Chao et al. (2000), and Corradi and Swanson (2000b). Other important papers in the burgeoning literature on predictive ability tests include Diebold and Mariano (1995), West (1996), Harvey et al. (1997), West and McCracken (1998), Clark (2000), Corradi and Swanson (2000a) and White (2000).

In addition to straightforward evaluation of the finite sample properties of various procedures, we address a number of additional empirical issues, such as:

1. Which model selection criteria perform best for choosing lags in finite sample scenarios, when the objective is to construct classical hypothesis tests of noncausality?
2. What is the relative impact of over- versus under-specification of the true lag order and/or the cointegration rank?
3. DGPs used include simple VAR(1) models and more complex VAR(3) models [possibly with moving average (MA) errors].

This allows us to examine not only which tests perform well in the context of simple models, but also in the context of more complex models.
We ask the question: is it true a procedure which performs "best" in the context of a simple VAR(1) model may be dominated by another test when more complex DGPs describe the data? Further, we are able to examine the rate of increase in test size distortion as model complexity, as well as sample size, increases.

The rest of the chapter is organized as follows: In Section 2, a variety of different procedures for ascertaining causal ordering are summarized and discussed. In Section 3, our experimental setup is discussed. Monte Carlo results are summarized in Section 4, and Section 5 concludes and offers a number of recommendation. Throughout, \( \rightarrow_d \) signifies convergence in distribution.

## 2 SUMMARY OF ALTERNATIVE TESTS AND MODEL SELECTION PROCEDURES

For simplicity, we start by assuming that an \( n \)-vector time series, \( y_t \), is generated by a \( K \)th order VAR model:

\[
y_t = J(L)y_{t-1} + u_t, \quad t = -K + 1, \ldots, T
\]

where \( J(L) = \sum_{i=1}^{K} J_i L^{-i} \). For simplicity of exposition, we leave the discussion of additional deterministic components in Eq. (1) until later. Assume also that

A.1. \( u_t = (u_{1t}, \ldots, u_{nt}) \) is an i.i.d. sequence of \( n \)-dimensional random vectors with mean zero and covariance matrix \( \Sigma_u > 0 \) such that \( E|u_t|^2+\delta < \infty \) infinity for some \( \delta > 0 \).

As we consider both integrated (integration of order 0 or higher) and cointegrated [by cointegrated we always refer to linear combinations of \( I(1) \) variables, which are \( I(0) \) as discussed in Engle and Granger (1987)] economic variables, a number of alternative assumptions on the \( J \) matrices are in order. The cases below where error correction models are estimated, we assumed that:

A.2. \( |I_0 - J(z)z| = 0 = > |z| > 1 \) or \( z = 1 \).

A.3. \( J(1) - I_0 = \Gamma A \) where \( \Gamma \) and \( A \) are \( n \times r \) matrices of full column rank \( r, 0 \leq r \leq n-1 \).

A.4. \( \Gamma_\rho (J^*(1) - I_0)A_\rho \) is non-singular, where \( \Gamma_\rho \) and \( A_\rho \) are \( n \times (n-r) \) matrices of full column rank such that \( \Gamma_\rho \Gamma = 0 = A_\rho A \).

Assumptions A.2–A.4 are standard, are taken from Toda and Phillips (1994), and ensure that Eq. (1) can be written in error-correction form:

\[
\Delta y_t = J^*(L)\Delta y_{t-1} + \Gamma A^* y_{t-1} + u_t
\]

where \( J^*(L) = \sum_{i=1}^{K-1} J_i L^{i-1} \). Throughout, we assume also that the covariance matrix of the stationary component in the system is
greater than zero. Finally, in the following discussion assume that we are interested in testing the hypothesis:

\[ H_0 : f(B) = 0 \quad \text{against} \quad H_A : f(B) \neq 0 \]

where \( f(\cdot) \) is an \( m \)-vector valued function satisfying standard assumptions. To simplify, we focus on the case where \( f(\cdot) = C\beta \), and \( C \) is a standard \( R \times (n^2 \times K) \) linear restriction matrix of rank \( R \), where \( R \) is the number of restrictions [see Lütkepohl (1991) for further details].

### 2.1 Wald-Type Tests

#### 2.1.1 All Variables Are Stationary

Assuming that \( y_t \) is generated by a stationary, stable VAR(\( K \)) process [see Lütkepohl (1991) for one definition of stability, e.g., assume A.1 and A.2 where \(|z| > 1\)], a frequently used version of the Wald test has

\[
\lambda_{ST} = C\hat{\beta}(CZZ')^{-1} \otimes \hat{\Sigma}_uC' \otimes C\hat{\beta} \xrightarrow{d} \chi^2_R
\]

where \( \hat{\Sigma}_u \) is a consistent estimator of the covariance matrix of \( u \) (e.g., \( \hat{\Sigma}_u = T^{-1}\hat{uu}' \) or \( \hat{\Sigma}_u = (T - nK)^{-1}\hat{uu}' \)), and \( \hat{\beta} \) is the least squares estimate of \( \text{vec}(B) \) from levels regressions of the form:

\[ Y = BZ + U \]

where \( Y \equiv ((y_1, \ldots, y_T) \) is the \( K \times T \) matrix of data, \( B \equiv ((J_1, \ldots, J_K) \), \( Z_t \equiv ((y_{t-K+1}, \ldots, y_t) \), \( Z_t^T \), \( Z_t^T \), \( B \equiv (\text{vec}(B) \), and \( U \equiv ((u_1, \ldots, u_T) \) is an \( n \times T \) matrix of errors. Of course, it has been known for quite some time that the \( \chi^2_R \) distribution is not generally appropriate under the null hypothesis, when variables are integrated and/or cointegrated, so that in our experiments we do not expect particularly good results from this statistic when data are generated under assumptions of difference stationarity and/or cointegration. Indeed, the focus of our Monte Carlo experiments is on cases where variables are difference stationary and/or cointegrated. Finite sample properties of causal order tests under stationary data are well known, and are thus not included here. However, full results for stationary variable cases are available in an earlier working paper version (Swanson et al., 1996). In addition, the reader is referred to Geweke et al. (1983).

#### 2.1.2 All Variables Are Difference Stationary

Assuming that A.1–A.4 hold and that there are \( n \) unit roots in the system (in which case we set \( A_P = I_n = \Gamma_P \), and \( J(1) = I_n \) it is reasonable to use \( \lambda_{ST} \) to test the null hypothesis, but with differenced data, and \( K - 1 \) lags. We will call this statistic \( \lambda_{DS} \).
2.1.3 Variables Are Integrated of any Order, and Possibly Cointegrated

Assuming that A.1 holds, consider the case where A.2–A.4 may hold. Alternatively, given A.1, assume that each element of \( \{y_t\} \) may be integrated of any order [e.g., \( I(0) \) or \( I(1) \)], and that any combination of variables in \( \{y_t\} \) may be cointegrated. Further, assume that we do not know, are unable to determine, or do not care what the orders of integratedness and/or cointegratedness are. Interestingly, in this case, Toda and Yamamoto (1995), and in independent work Dolado and Lütkepohl (1996), show that standard Wald tests can be applied to the coefficient matrices up to the correct lag order, when the system in levels [Eq. (1)] is artificially augmented by including at least as many extra lags of each variable as the highest order of integration of any of the variables in the system. (This test is hereafter referred to as the TYDL test.) In particular, it turns out that asymptotic chi-squared tests of causality restrictions can be applied to the submatrix of coefficients up to the correct lag order. The reason for this property is that all nonstandard asymptotics are essentially confined to the coefficient matrices beyond the correct lag order, and standard asymptotics apply with respect to the coefficient matrices up to the correct lag order. [Choi (1993) uses similar arguments to show that standard asymptotics can be applied to unit root tests which are suitably augmented.] Overall, the method of including surplus lags provides a very tractable, and hence appealing means with which to deal with many of the nonstandard distribution problems that characterize so many economic variables. However, it should be pointed out that the method is inefficient in the sense that extra coefficient matrices are estimated, so that the methods’ power may be affected in finite samples. Indeed, it remains to see what the local power properties of such tests are. As the test is a standard “levels regression” type Wald test, it is clearly very easy to apply, and has the form:

\[
\lambda_{LW} = C\hat{B}((ZQZ')^{-1} \otimes \hat{\Sigma}_e)C^{-1}C\hat{B} \rightarrow_d \chi^2_R
\]

where all regressions are the same standard least squares regressions estimated using levels data (i.e., as used to construct \( \lambda_{ST} \)), except that extra lags of the regressors are added as surplus explanatory variables. That is,

\[
Y = \hat{B}^*Z + \hat{\Psi}W + \hat{\epsilon}
\]

where \( \hat{\Sigma}_e = T^{-1}\hat{\epsilon}\hat{\epsilon}' \). Note that \( \hat{B}^* \), which is estimated using Eq. (3), is not the same as the estimator of \( B \) used in \( \lambda_{ST} \) and \( \lambda_{DS} \), as \( W \) extra variables are added to the standard levels model, where \( W_t \equiv (y_{t-K}, \ldots, y_{t-K-d+1})' \), and \( W \equiv (W_0, \ldots, W_{T-1}); \hat{\Psi} \) is the coefficient matrix associated with these extra lags, and \( d \) is the highest order of integration of any of the variables.
in the system. The matrix $Q$ is defined in the standard way as: 

$$Q = Q_t - Q_t W(WQ_t W)^{-1} WQ_t,$$

where $Q_t = I_T - S(S')^{-1} S$, where $S$ is a matrix that contains constant and/or deterministic trend variables. In our case, we have set $S$ to zero. However, if $S$ is nonzero, then the variables can be added in the obvious way as additional regressors in Eq. (3), without affecting the distributional result under the null hypothesis.

One feature of the TYDL method that deserves further mention is that any form of Granger causal relation is captured by the test, regardless of whether it is “long run”, in the sense that error-correction terms are constructed using the variable(s) whose causal effect is being examined, or “short run”, in the sense that lagged differences of the relevant variable(s) appear in the error-correction model, or both. Alternatively, the sequential Wald tests discussed below have the feature that both “short-run noncausality” and “long-run noncausality” are sequentially tested for.

However, as we will see below, if we incorrectly specify a difference stationary model (when the true model is cointegrated), then the estimated coefficient matrices become so “contaminated” that standard Wald tests (i.e., $\lambda_{pS}$) suffer from severe size distortion, often rendering the results of tests of short-run noncausality” meaningless. Thus, it is not sufficient simply to model differenced data if all that we are interested in is “short-run noncausality”. The reader is referred to Granger and Lin (1995) and Toda and Phillips (1994) for further discussion of short- and long-run causality.

### 2.1.4 Variables Are Integrated, and Possibly Cointegrated

Toda and Phillips (1993) provide a theory for Wald tests of Granger noncausality in levels vector autoregressions that allow for stochastic and deterministic trends as well as arbitrary degrees of cointegration. The theory extends earlier work by Sims et al. (1990) on trivariate VARs and suggests a tractable strategy for implementing a variety of Wald-type tests. An important feature of this approach is that tests are based on estimated versions of Eq. (2), the error correction model. Thus, while potentially offering gains in efficiency when estimating the system (relative to the TYDL approach of estimating the system in levels) the sequential Wald test adds a layer of estimation to the problem, in the sense that cointegrating ranks must be first estimated. Nevertheless, as shown in Toda and Phillips (1994), the sequential Wald tests appear to have good size properties, for a number of alternative specifications. To be more precise about the null hypothesis of interest for sequential Wald tests, we here consider variations of the joint null:

$$H_0: J_{i,13} = \cdots = J_{K-1,13} = 0 \quad \text{and} \quad \Gamma_1 A_3' = 0$$
where $J'_{13}(L) = \sum_{i=1}^{K-1} L^{i-1}$ is the $n_1 \times n_3$ upper right submatrix of $J'(L)$, $\Gamma_1$ is the first $n_1$ rows of the loading coefficient matrix $\Gamma$, and $y_t = (y_{1t}, y_{2t}, y_{3t})'$ has been partitioned into three vectors of length $n_1$, $n_2$, and $n_3$, respectively.

Before discussing the sequential Wald tests any further, it is worth reiterating an important point that is made (and elaborated on) in Toda and Phillips (1994) [and also to some extent in Sims et al. (1990)]. In the case of cointegrated economic variables, standard levels regression based causality tests are valid asymptotically as $\chi^2$ criteria when there is sufficient cointegration among the variables whose causal effects are being tested. However, the rank condition associated with this criterion suffers from simultaneous equations bias when levels VARs are estimated, suggesting that there is no valid statistical basis in VARs for testing if the condition holds. Furthermore, when the rank condition fails, there is no valid statistical basis for mounting a standard Wald test of causality. One interpretation of this point is simply that if we are using the “correct” specification of Eq. (1) (e.g., we have specified the correct number of lags) and there is cointegration among the variables, then standard Wald tests on levels are not directly feasible, and sequential Wald tests that are based on error-correction models are. Of final note is that, if there is no cointegration, the standard levels regression-based Wald statistic (i.e., $\lambda_{ST}$) for causality has a nuisance parameter-free nonstandard limit distribution, and so can conceivably be used. These points are explained in some detail in Toda and Phillips (1994). We now state the three sequential procedures, and the testable hypotheses associated with them. The actual test statistics are given in Section 2 of Toda and Phillips (1994). For simplicity, assume that $\Gamma_1$ and $A_3$ are $r$, or $\hat{r}$-dimensional row vectors, and that $K > 1$ (with obvious modifications otherwise).

Then, the tests are based on the null hypotheses:

- $H_1': J'_{1,13} = \cdots = J'_{K-1,13} = 0$
- $H_2': J'_{1,13} = \cdots = J'_{K-1,13} = 0$
- $H_3': \Gamma_1 = 0$
- $H_4': A_3 = 0$
- $H_4': \Gamma_3 A_3 = 0$

and as above:

- $H_0^*: J'_{1,13} = \cdots = J'_{K-1,13} = 0$ and $\Gamma_1 A_3 = 0$

The sequential testing procedures are:

P1. Test $H_1'$. If $H_1'$ is rejected, test $H_0^*$ using a $\chi^2_{n_1K}$ critical value. Otherwise, test $H_1'$.
P2. Test $H_2$. If $H_2$ is rejected, test $H_0$ using a $\chi^2_{n,k}$ critical value. Otherwise, test $H'_1$.

P3. Test $H_3$. If $H_3$ is rejected, reject the null hypothesis of non-causality. Otherwise, test $H'_1$ and $H'_2$. If both are rejected, test $H'_0$ if $r > 1$, or reject the null if $r = 1$. Otherwise, accept the null of noncausality. These tests are easy to apply once standard procedures have been used to estimate the rank of the cointegrating space, and the number of cointegrating vectors in the system, although it should be noted that exact control of the overall size of these causality tests is not feasible. In what follows, we will refer to the test statistics associated with P1–P3 as $\lambda_{P1}$, $\lambda_{P2}$, and $\lambda_{P3}$, respectively.

In closing, note that while the sequential Wald discussed above is essentially designed to deal with $I(1)$ variables, which may or may not be cointegrated [of order (1,1) using the notation of Engle and Granger (1987)], the TYDL method allows for any mixtures of $I(0)$, $I(1)$, $I(2)$, etc., variables, which may or may not be cointegrated. Finally, the FM–VAR test (see below) also requires that A.1–A.4 hold [e.g., $|L - J(L)L| = 0$ has roots on or outside the unit circle], with the modification that assumption A.3 can be weakened so that $0 \leq r \leq n$, allowing for the possibility that all of the variables are stationary, for example. Of course, one reason why the sequential Wald test has $r < n$ stems from the fact that it is assumed that the variables are pretested to determine their order of integration, and if all are found to be $I(0)$ then standard chi-squared asymptotics apply, and sequential Wald tests are not called for.

2.1.5 Variables Are Integrated, and Possibly Cointegrated II

Another type of Wald causality test, which is in keeping with the unrestricted levels VAR method of TYDL, is that based on fully modified vector autoregression (FM-VAR) (Phillips, 1995). Phillips proposes making corrections to standard OLS–VAR regression formulas, which account for serial correlation effects as well as for regressor “endogeneity” that results from the existence of cointegrating relationship(s). The endogeneity in the context of cointegrated models arises for the following reason. Assume that Eq. (1) can be partitioned into two subsystems (1) $y_{1t} = D_{1t}y_{2t-1} + u_{1t}$, and (2) $y_{2t} = y_{2t-1} + u_{2t}$. The standard procedure is to treat $y_{2t-1}$ as predetermined, and regard the model as a reduced form. However, even though $E(u_{1t}y_{2t-1}) = 0$, the sample covariance, $T^{-1} \sum_{t=1}^{T} u_{1t}y_{2t-1}^{'},$ does not converge to zero, and instead converges to a function of Brownian motions, which are generally correlated in the limit.

The FM–VAR method is largely a generalization of work by Phillips and Hansen (1990), and has the feature that advance knowledge of which variables are stationary, nonstationary, and/or cointegrated is not required.
to achieve consistency of the fully modified \( V \)AR estimator. Given a
kernel condition (Phillips, 1995, p. 1031), a bandwidth expansion rate
1044), which is essentially the same as assumptions A.1–A.4 above, the
relevant Wald-type test statistic is

\[
\lambda_{FM} = C\hat{\beta}^T(C(XX')^{-1} \otimes \hat{\Sigma})C^{-1}C\hat{\beta}^T
\]

where \( X \) is a matrix of observations with the same informational content
of \( Z \) above, except that it is comprised of \( K-1 \) differences of the
variables, as well as the lagged level of the variables; \( \hat{\Sigma} \) is the usual OLS
estimate of the error covariance matrix from a regression of \( Y \) on \( X \). The
test used here is actually distributed as a weighted sum of \( \chi^2 \) values, whose
weights are difficult to estimate. However, as pointed out by Phillips
(1995), an asymptotically conservative test can be constructed by using the
\( \chi^2 \) distribution as the limiting distribution for \( \lambda_{FM} \). We adopt that
approach in the sequel. Although the reader is referred to Phillips (1995)
for further details, note that the FM–\( V \)AR estimator, \( \hat{\beta}^T = \text{vec}(\hat{\beta}^T) \), is

\[
\hat{F}^+ = [XY_D'YY_{-1} - \hat{\Omega}_xy\hat{\Omega}_{yy}^{-1}(\Delta Y_{-1}Y_{-1} - T\hat{\Delta}_{xy}\Delta_{yy}')(XX')^{-1}
\]

where \( X_D \) is the same as \( X \), except that the levels variables are removed, \( Y_{-1} \)
is \( Y \) lagged one period, and the \( \Omega \) and \( \Delta \) estimates are long-run covariance
matrix estimates, and one-sided long-run covariance matrix estimates
calculated in the usual way as the weighted sum of estimated auto-
covariance terms, using some appropriate kernel function as the weighting
scheme [see Phillips (1995) for details]. Deterministic components can also
be added to the FM–\( V \)AR regression model by augmenting the matrix of
explanatory data used in the construction of \( \lambda_{FM} \) in the usual way.

2.2 Model Selection-Type Causality Procedures

The next two procedures that can be used to examine causal ordering are
not classical, in the sense that critical values are not used. In practice,
these types of procedures involve choosing a “best” model. If that “best”
model contains the variable(s) whose causal effect is being examined, then
we “choose” the model that is not consistent with a null hypothesis of
noncausality. The “best” model may be a single equation, or a model
consisting of many equations, depending on how many variables are in
\( y_1 \), when we are testing whether \( y_3 \) is noncausal for \( y_1 \). One advantage of
model selection, e.g., based on complexity penalized likelihood criteria, is
that the approach does not require specification of a correct model for its
valid application. Another desirable feature is that the probability of
selecting the truly best model approaches one as the sample size increases,
if the model selection approach is properly designed. This is contrary to the standard practice of fixing a test size, and rejecting the null hypothesis at that fixed size, regardless of sample size. An interesting result of our simulations, which is in part a consequence of these features (see below), is that the empirical rejection frequencies based on both Akaike and Schwarz information criteria (AIC and SIC, respectively) become vanishingly small (in many cases) for sample sizes of only 300 observations, regardless as to whether the “true” model is among those in the “selection” set.

As mentioned above, Granger causality tests have a natural interpretation as tests of predictive ability. In this sense, it may seem natural to use the results of causality tests when constructing forecasting models. However, it is commonly argued that any good forecasting model should be “tested” using some sort of ex ante forecasting experiment before being implemented in any practical setting. Along these lines, a reasonable alternative* procedure for checking causal ordering involves comparing competing forecasting models, both with and without the variable whose causal effect is being examined. These “competing” models could then be subjected to ex ante forecasting analysis and the “best” model chosen using some sort of out-of-sample loss function (e.g., mean square prediction error), or by constructing an out-of-sample predictive ability test. If the “best” model contains the variable of interest, noncausality is rejected. In general, there are many model selection criteria to choose from, including forecast error summary measures, directional forecast accuracy measures, and profit measures, for example. Choice of criterion is naturally dependent on the objective function of the practitioner. Diebold and Lopez (1995) and Swanson and White (1997a,b) discuss these and related issues, and contain other related references.

2.3 Complexity-Based Likelihood Criteria

Granger et al. (1995) suggest that although standard hypothesis testing has a role to play in terms of testing economic theories, it is more difficult to justify using standard hypothesis tests for choosing between two competing models. One reason for their concern is that one model must be selected as the null, and this model is often the more parsimonious. However, it is often difficult to distinguish between the two models (because of multicollinearity, near-identification, etc.), so that the null hypothesis may be unfairly favored. For example, it is far from clear that pretest significance levels of 5 and 1% are optimal, as pointed out by

*In our experiments, we use the predictive ability test due to Diebold and Mariano (1995) to “compare” our alternative models.
Fomby and Guilkey (1978) in the context of Durbin–Watson tests. The use of model selection criteria neatly avoids related sticky issues associated with how to test theories and how arbitrarily to choose significance levels. Further, as discussed above, model selection procedure have a number of optimal properties is used to “select” among competing models, and seem a natural alternative to standard hypothesis testing when considering issues of Granger causality and/or predictive ability.

Sin and White (1995) consider the use of penalized likelihood criteria for selecting models of dependent processes. In the context of strictly nested, overlapping or non-nested, linear or nonlinear, and correctly specified or mis-specified models they provide sufficient conditions on the penalty to ensure that the model selected attains the lower average Kullback–Leibler information criterion, with probability (approaching) one. Further, as special cases, their results describe the AIC and SIC.

Our experiments into the use of model selection criteria focus on their ability to identify “correctly” and unravel the causal relationships among stationary, integrated, and cointegrated time series variables. If it turns out that these model selection criteria perform well, then we have direct evidence that model selection criteria offer a valid alternative to standard hypothesis testing in the context of uncovering causal relationships. Further, as model selection criteria are very easy to calculate, regardless of the properties of the data, evidence of the usefulness of AIC and SIC might be of some interest to empirical economists. The two model selection criteria which we examine below are the AIC (Akaike, 1973, 1974):

\[ AIC = T \log |\hat{\Sigma}| + 2f \]

and the SIC (Rissanen, 1978; Schwarz, 1978):

\[ SIC = T \log |\hat{\Sigma}| + f \log (T) \]

where we define \( f \) to be the total number of parameters in the system [if we are measuring the causal effect of some variable(s) on a group of more than one other variable], or \( f \) is the number of parameters in the single equation of interest (when the group being examined is a single variable). Similarly, \( \hat{\Sigma} \) is some standard estimate of the error covariance matrix, which is scalar if only one equation in the system is being examined.

2.4 Ex Ante Forecast Comparison

Although AIC and SIC may prove to be valuable when testing the null hypothesis of non causality, note that they are calculated “in sample,” as are the standard Wald-type tests discussed in Section 2.1. However, if, as is often the case, the goal of our analysis is to construct an “optimal” forecasting model, then causality might be better tested for in an ex ante
rather than *ex post* setting, as discussed above. After all, such procedures would allow us to tackle directly the issue of predictive ability. In addition to those papers cited above, Engle and Brown (1986), Fair and Shiller (1990), and Swanson (1998) discuss *ex ante* model selection. However, these papers, as well as Diebold and Mariano (1995), West (1996), Swanson (1998), White (2000), and Corradi and Swanson (2000a) consider the application of predictive accuracy tests, which are not shown to be valid in the context of nested models (i.e., when forming Granger causality-type predictive ability tests). For tests that are valid in the context of nested models, see Clark and McCracken (1999), McCracken (1999), Chao et al. (2000), and Corradi and Swanson (2000b). The approach that we adopt involves constructing sequences of real-time one-step ahead forecasts of the variable of interest using a variety of models (levels VARs, differences VARs, and VECs) as well as a variety of different lag structures, cointegration assumptions, etc. The resulting sequences of forecasts are then used to construct three measures of out-of-sample forecast performance. The measures are mean squared forecast error ($MSE = \sum_{t=1}^{T} \hat{e}_t^2 / T$), and the well known related statistics: mean absolute forecast error deviation (MAD) and mean absolute forecast percentage error (MAPE), where $T^*$ is the out-of-sample period (which here varies with sample size), while $\hat{f}_t$ is the appropriate forecast error associated with the particular econometric model in question. The “best” model is not selected based on direct examination of MSE, MAD, and MAPE criteria, however. Rather, we construct predictive accuracy tests described in Diebold and Mariano (1995), and implemented in Swanson and White (1997a,b). These test statistics are constructed by first forming the mean ($d$) of a given loss differential series, $d_t$ (e.g., for MSE we construct $d_t = \hat{e}_t^2, - \hat{e}_{t,t}^2$, where the subscript 1 corresponds to the null model, and subscript 2 denotes the same model as subscript 1, but with additional regressors corresponding to the variable whose causal effect is being examined). The loss differential test statistic is then constructed by dividing ($d$) by an estimate of the standard error of $d_t$, say $\hat{s}_d$. We use the parametric covariance matrix estimation procedure of Den Haan and Levin (1995) to estimate $\hat{s}_d$, in part because the method is very easy to implement, and in part because these authors show that their procedure compares favorably with a number of nonparametric covariance matrix estimation techniques. Further details of these tests are given in Swanson and White (1995, 1997). McCracken (1999) gives critical values that are valid when Diebold–Mariano tests are used to compare nested alternatives.

In our simulations we are able to implement a truly *ex ante* forecasting procedure. However, as is well known, the use of real-time forecasting methods becomes suspect when the actual economic data that
are used have been subjected to two-sided moving average filtering, periodic rebasing, and periodic revision, for example. Thus, we stress that these and related problems may make valid implementation of these types of tests quite difficult in practice. In Section 3 we discuss the design of a number of Monte Carlo experiments that were used to examine the finite sample properties of the above tests.

3 EXPERIMENTAL DESIGN

Our Monte Carlo experiments are all based on 5000 simulations, and start with various difference stationary and cointegrated versions of the following two VAR models:

\[ y_t = J_1 y_{t-1} + u_t, \quad t = 0, \ldots, T \]  
\[ y_t = J_1 y_{t-1} + J_2 y_{t-2} + J_3 y_{t-3} + u_t, \quad t = -2, \ldots, T \]

We use a VAR(1) and a VAR(3) model because it is well known that causality tests often result in quite different conclusions, depending on how many lags are used to estimate a model. Clearly, one reason why this problem arises is that models which are misspecified in the sense that surplus lags are included, may suffer decreased test power. However, another possibility is that increasing the number of lags estimated results in what we will call "least squares confusion" (i.e., the tendency for least squares to yield poor parameter estimates), even if the "true" data are generated by a process with many lags. In practice, this type of situation may arise because of multicollinearity, and would probably manifest itself in the form of an increased incidence of rejecting the null hypothesis of Granger noncausality, even when the null is true. If this is the case, then we would expect experimental results based solely on VAR(1) processes to be conservative relative to those based on higher order VARs. As empirical models often contain many lags, "least squares confusion" may be damaging in the sense that we may later construct tests under the assumption that their size is close to the nominal test size, when in actuality the size is much higher than the nominal size. Further, the tests that perform "best" might change as we increase the number of lags in the system. In fact, we will see later that these issues are relevant, and have a non-negligible impact on our findings.

Our data generating processes can be summarized as follows:

Cointegrated models (cointegrating rank = 1; \( I_3 - \sum J_i = \Gamma A' \))

\[
(VAR1 - C1) \quad J_1 = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ -1.0 & 1.0 & 1.0 \end{bmatrix} \quad \Gamma = \begin{bmatrix} 0.5 \\ 0.0 \\ 1.0 \end{bmatrix} \quad A = \begin{bmatrix} 1.0 \\ -1.0 \\ 0.0 \end{bmatrix}
\]
It should perhaps be noted that the above parameterizations are ad hoc in the sense that they are not calibrated by examining actual economic data, for example. In addition, economic data at most frequencies are usually modelled by including more than three lags. In this sense, the above DGPs are too simplistic. For these reasons, the Monte Carlo results reported below are meant only to be a guide to the sorts of relationships among the various causality tests, which might be observed in practice. In addition, the following structure is placed on the errors of the above DGPs, where \( u_t = \eta_t - \theta \eta_{t-1}, \) and \( \eta_t \) is a unit normal random vector:

\[
(MA1) \theta = 0
\]

\[
(MA2) \theta = \begin{bmatrix}
0.5 & 0.0 & 0.0 \\
0.0 & 0.5 & 0.0 \\
1.0 & 0.0 & 0.5
\end{bmatrix}
\]
These values of $\theta$ are the same as those used by Toda and Phillips (1994). As discussed in Toda and Phillips (1994), even though (MA2) appears to be inconsistent with A.1, $u_t$ is an invertible MA process so that VAR(1) and VAR(3) can be rewritten as infinite order VARs. As the higher lags are exponentially decaying in the parameters, the actual number of lags needed for a “reasonable” approximation is low. Further, it will be interesting to see how well our various statistics perform when faced with inherently mis-specified models (in the MA1 models, the true specification is always included within the set of models that are fit). Finally, the addition of MA2 to our experimental setup ensures that some of our models [i.e., the VAR(3) models] should ideally contain more than three lags when estimated.

In most cases we are testing the null hypothesis that $y_3$ is Granger noncausal for $y_1$ (we use the notation $y_3 \not\rightarrow y_1$ for cases such as this), and thus we focus somewhat on the empirical size of the seven tests. The exception is CI2, in which case we test the null hypothesis that $y_1 \not\rightarrow y_2$. In all experiments, whenever a test of the null that $y_j \not\rightarrow y_j$ is carried out, we also perform a test of the null that $y_j \not\rightarrow y_i$. This allows us to tabulate power at the same time that we tabulate empirical size. It should be noted that in the case of our model selection procedures, “empirical size” is not the appropriate measure. Instead, we report the rejection frequency of the correct model in favor of the “incorrect” model. As the tests that we examine all have good power properties (albeit rejection frequencies under alternative hypotheses are not “size adjusted”), these findings leave little to choose between the tests. Not surprisingly, then, may analyses related to ours focus on empirical size. We follow suit, and report only empirical size (and its analog for our model selection procedures). Complete finding are available on request from the authors.

In each of the experiments we use 100, 200, and 300 observations. Assuming that the true number of lags in each case is $p$, then the number of lags used in each estimation is set at \{ $p - 1$, $p$, $p + 1$, $p_{AIC}$, $p_{SIC}$ \}, where $p_{AIC}$ and $p_{SIC}$ use the number of lags selected by implementation of the AIC and SIC model selection criteria, where the model selection criteria are constructed using estimates of levels VARs. Also, all lags are reported in terms of a VAR in levels, so that results for difference VAR Wald-type tests ($\lambda_{DS}$) for one lag are not given. This is because a level VAR with one lag translates to a difference VAR with no lags.

In all experiments we calculate many difference stationary, stationary, and VEC systems, regardless of the true DGP. This is because the seven procedures considered variously involve estimation of all three types of VARs. For example, a cointegration model is estimated, and sequential Wald tests are constructed, even when the true data-generating
process is stationary. One reason for an extensive analysis of this type is that absolute size distortions associated with each test type are not necessarily as important as the size distortion of one test relative to the size distortion of other tests, regardless as to whether either test is theoretically valid. As an example of why this issue is perhaps important, note that Gonzalo and Lee (1996) show that, while it is well known that augmented Dickey–Fuller tests suffer from low power when processes have near unit roots, and suffer from size distortions in other cases, it is less widely acknowledged that standard \( t \)-statistics also have similar power and size problems, even when the null hypothesis is that \( \beta = 0.5 \) and the true process is a stationary AR(1) with slope coefficient \( \beta \). Extending this argument, it should be of interest to determine whether size and power distortions associated with using standard \( \chi^2 \) asymptotics when they are not valid are large enough to warrant the use of more complex (albeit asymptotically valid) testing procedures which themselves may suffer from size and power distortions of other types. In particular, even though asymptotically valid, the finite sample performance of some of the more complex test statistics considered here is not certain.

We also used \( F \)-test versions of all Wald-type tests, as it is well known that these often have better size properties in finite samples. To be more specific, we construct \( F = \frac{\lambda}{R} \), where \( R \) is the number of restrictions. These statistics are assumed to be distributed as \( F \), with \( T - \hat{f} \) degrees of freedom in the denominator, where \( \hat{f} \) is (1) the number of parameters in the equation being examined, and \( f \) is also defined as (2) the number of parameters estimated in the entire system. Interestingly, the properties of these versions of the Wald tests were qualitatively the same as when the \( \chi^2 \) versions of the tests were used. For this reason, these results are not reported in the tables, although complete results are available in an earlier working paper version of this study.

Whenever VEC models are estimated, we use the maximum likelihood procedures developed by Johansen (1988, 1991) and Johansen and Juselius (1990). In particular, assume that the true cointegrating rank is \( r \). Then, the cointegrating ranks used are usually \( \{r - 1, r, r + 1, r_{JO1}, r_{JO2}\} \).

The last two are rank estimates based on the well known likelihood ratio trace statistic using a 5% size, and with and without a constant in the estimated system, respectively. Results based on the related maximum statistic are not reported for the sake of brevity. The exception to this rule are cases where the true cointegrating rank is zero. In these cases, we use \( \{r + 1, r, r + 2, r_{JO1}, r_{JO2}\} \), where \( r_{JO1} \) and \( r_{JO2} \) may be zero, in which case VEC models are not estimated, and corresponding sequential Wald test results are not reported. With the exception of the test statistic based on
fully modified VAR estimates, all other estimations done in the experiments are multivariate least squares.

Overall, we calculate all Wald-type tests whenever possible. However, it should be noted that there are some cases where Wald-type tests are not applicable. As mentioned above, difference stationary versions of the standard Wald test are not calculated when the lag is one (as a one lag VAR in levels is a zero lag VAR in differences). Also, there are some lag structures and cointegrating rank cases (e.g., \( r = n \)) for which the sequential Wald-type tests are not applicable, and are not calculated. [The reader is referred to Toda and Phillips (1994) for more details.] In such cases, the reported simulation results are calculated using possibly fewer than 5,000 simulation trials. Instances where this situation arises are reported on in the next section.

The AIC and SIC type tests are implemented as follows. For all lags, cointegrating ranks, and alternative models (i.e., levels VAR, differences VAR, an VEC) AIC and SIC statistics are calculated. In particular, the statistics are calculated for versions of the models both with and without the variable(s) whose causal effect is being examined. As a summary test, the AIC and SIC best models are chosen. If the “best” model contains any of the relevant variable(s), then the null hypothesis of Granger noncausality is rejected. We also track with type of model is chosen as “best” in each simulation. These types of tests are individually reported on for each lag and each model (e.g., cointegrating model with cointegrating space rank of unity), as well as across all models. The latter reporting style yields what we will call “global” tests, while the former is more in keeping with a comparison of only “two” alternative models when testing the null hypothesis of noncausality.

Ex ante forecasting tests are carried out the same way as the AIC and SIC type tests, with two exceptions. First, the statistics calculated are MSE, MAD, and MAPE. Second the statistics are constructed from sequences of one-step ahead forecast errors, and these forecast errors are in turn used to construct predictive accuracy test statistics (as discussed above). The procedure is implemented in the following way. A starting sample of 0.7 \( T \) observations is used to construct a one-step ahead forecast, using all different model types, lags, etc., and for \( T = 100, 200, \) or 300. From this, the first forecast error is saved. One more observation is then added to the sample, and another set of models are estimated, forecasts are constructed, and forecast errors are saved. This procedure continuous until the entire sample is exhausted, and the resultant \( ex \ ante \) sequence of \((0.3T) - 1\) forecast errors is used to construct the model selection criteria. It should be noted that one case arises where the actual number of forecast errors used to construct the model selection criteria is not \((0.3T) - 1\). For the VEC models that are estimated using cointegrating
rank, \( r = r_{JO1} \) or \( r = r_{JO2} \), there may be some simulations where the estimated cointegrating rank is 0 or \( n \) (in which case the model is not estimated). This situation arises most frequently when the actual DGPs are stationary or difference stationary VARs. In these cases, we will report the average number of observations used in the construction of the selection criteria. In addition, it should be noted that, due to possible model misspecification (see, e.g., Corradi and Swanson, 2000b), finite sample results may not be the same for the one-step ahead versus multistep ahead approach to \textit{ex ante} forecasting implementations of Granger causality analysis.

4 RESULTS

A summary of our experimental findings is given in Tables 1–15. Complete experimental results are contained in an earlier working paper version of this chapter (i.e., see Swanson et al. and Pisu, 1996). Our findings from the Monte Carlo experiments are fairly clear cut, and can be summarized as follows:

1. In all cases (i.e., regardless of lag and model specification) the Toda and Yamamoto type Wald tests have empirical sizes close to nominal (nominal size is set at 5% throughout).

2. Sequential Wald tests, and standard Wald tests based on the use of differenced data only perform the best when correctly specified models (i.e., correct lags, cointegrating rank, data transformation) are estimated.

3. SIC type approaches, and to a lesser degree those based on the AIC perform quite well, even when the DGP has moving average (MA) error components (i.e., when the true lag order of the model is infinite and is hence only approximated using AIC and SIC measures). In many cases and for samples of moderate and large size (200 and 300 observations), the SIC-type causality test has rejection frequency approaching zero very rapidly, and power close to unity.

4. As expected, test performance (with the exception of all of our model selection type causality “tests”) worsens substantially when model complexity is increased (i.e., going from one to three lags and from no MA to an MA component in the true DGP).

5. Lag selection based on the AIC and SIC does very well. In particular, the lag structure chosen using the SIC almost always yields the “best” results when the DGP is a VAR(1) with no MA component. However, for other DGPs, the lag structure given by the AIC generally results in better test performance.

6. \textit{Ex ante} forecasting type approaches tend to perform well [see Chao et al. (2000) and McCracken (1999) for further evidence]. In particular, when
<table>
<thead>
<tr>
<th>DGP/sample</th>
<th>Lag</th>
<th>(CI-1)</th>
<th>(CI-2)</th>
<th>(DSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{\lambda}_{ST}$</td>
<td>$\hat{\lambda}_{DS}$</td>
<td>$\hat{\lambda}_{LW}$</td>
</tr>
<tr>
<td>VAR(1)-(MA1)</td>
<td>1</td>
<td>0.167</td>
<td>0.065</td>
<td>0.146</td>
</tr>
<tr>
<td>$T = 100$</td>
<td>2</td>
<td>0.142</td>
<td>0.064</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.175</td>
<td>0.064</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.177</td>
<td>0.065</td>
<td>0.059</td>
</tr>
<tr>
<td>$T = 200$</td>
<td>1</td>
<td>0.147</td>
<td>0.068</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.156</td>
<td>0.067</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.147</td>
<td>0.068</td>
<td>0.057</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>1</td>
<td>0.162</td>
<td>0.048</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.195</td>
<td>0.047</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.147</td>
<td>0.048</td>
<td>0.051</td>
</tr>
<tr>
<td>VAR(1)-(MA2)</td>
<td>1</td>
<td>0.244</td>
<td>0.154</td>
<td>0.407</td>
</tr>
<tr>
<td>$T = 100$</td>
<td>2</td>
<td>0.217</td>
<td>0.108</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.215</td>
<td>0.114</td>
<td>0.207</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.238</td>
<td>0.074</td>
<td>0.207</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>DGP/sample</th>
<th>Lag</th>
<th>( \lambda_{ST} )</th>
<th>( \lambda_{DS} )</th>
<th>( \lambda_{LW} )</th>
<th>( \lambda_{FM} )</th>
<th>( \lambda_{ST} )</th>
<th>( \lambda_{DS} )</th>
<th>( \lambda_{LW} )</th>
<th>( \lambda_{FM} )</th>
<th>( \lambda_{ST} )</th>
<th>( \lambda_{DS} )</th>
<th>( \lambda_{LW} )</th>
<th>( \lambda_{FM} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T = 200 )</td>
<td>2</td>
<td>0.274</td>
<td>1.00</td>
<td>0.106</td>
<td>0.861</td>
<td>0.217</td>
<td>1.00</td>
<td>0.063</td>
<td>0.996</td>
<td>0.504</td>
<td>0.505</td>
<td>0.134</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.169</td>
<td>1.00</td>
<td>0.079</td>
<td>0.260</td>
<td>0.159</td>
<td>0.994</td>
<td>0.082</td>
<td>0.719</td>
<td>0.188</td>
<td>0.117</td>
<td>0.104</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.251</td>
<td>1.00</td>
<td>0.093</td>
<td>0.724</td>
<td>0.170</td>
<td>1.00</td>
<td>0.069</td>
<td>0.406</td>
<td>0.413</td>
<td>0.363</td>
<td>0.106</td>
<td>0.710</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.258</td>
<td>—</td>
<td>0.285</td>
<td>—</td>
<td>0.447</td>
<td>—</td>
<td>0.055</td>
<td>—</td>
<td>0.813</td>
<td>—</td>
<td>0.400</td>
<td>—</td>
</tr>
<tr>
<td>( T = 300 )</td>
<td>2</td>
<td>0.324</td>
<td>1.00</td>
<td>0.119</td>
<td>0.908</td>
<td>0.226</td>
<td>1.00</td>
<td>0.059</td>
<td>1.00</td>
<td>0.656</td>
<td>0.681</td>
<td>0.165</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.140</td>
<td>1.00</td>
<td>0.066</td>
<td>0.266</td>
<td>0.140</td>
<td>1.00</td>
<td>0.073</td>
<td>0.687</td>
<td>0.158</td>
<td>0.099</td>
<td>0.084</td>
<td>0.587</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.217</td>
<td>1.00</td>
<td>0.078</td>
<td>0.560</td>
<td>0.151</td>
<td>1.00</td>
<td>0.061</td>
<td>0.384</td>
<td>0.321</td>
<td>0.278</td>
<td>0.090</td>
<td>0.666</td>
</tr>
<tr>
<td>( VAR(3)-(MA1) )</td>
<td>2</td>
<td>0.185</td>
<td>0.530</td>
<td>0.075</td>
<td>0.554</td>
<td>0.162</td>
<td>0.236</td>
<td>0.074</td>
<td>0.605</td>
<td>0.371</td>
<td>0.099</td>
<td>0.086</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.139</td>
<td>0.874</td>
<td>0.085</td>
<td>0.229</td>
<td>0.132</td>
<td>0.276</td>
<td>0.088</td>
<td>0.165</td>
<td>0.145</td>
<td>0.066</td>
<td>0.103</td>
<td>0.408</td>
</tr>
<tr>
<td>( T = 100 )</td>
<td>4</td>
<td>0.147</td>
<td>0.895</td>
<td>0.100</td>
<td>0.276</td>
<td>0.137</td>
<td>0.256</td>
<td>0.100</td>
<td>0.175</td>
<td>0.150</td>
<td>0.077</td>
<td>0.123</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.170</td>
<td>0.797</td>
<td>0.089</td>
<td>0.308</td>
<td>0.162</td>
<td>0.280</td>
<td>0.093</td>
<td>0.261</td>
<td>0.171</td>
<td>0.089</td>
<td>0.114</td>
<td>0.406</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.192</td>
<td>0.547</td>
<td>0.076</td>
<td>0.536</td>
<td>0.168</td>
<td>0.243</td>
<td>0.075</td>
<td>0.575</td>
<td>0.145</td>
<td>0.066</td>
<td>0.103</td>
<td>0.408</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.169</td>
<td>0.821</td>
<td>0.061</td>
<td>0.565</td>
<td>0.160</td>
<td>0.399</td>
<td>0.060</td>
<td>0.814</td>
<td>0.382</td>
<td>0.106</td>
<td>0.070</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.110</td>
<td>0.996</td>
<td>0.064</td>
<td>0.198</td>
<td>0.113</td>
<td>0.466</td>
<td>0.067</td>
<td>0.143</td>
<td>0.132</td>
<td>0.057</td>
<td>0.076</td>
<td>0.473</td>
</tr>
<tr>
<td>( T = 200 )</td>
<td>4</td>
<td>0.107</td>
<td>0.997</td>
<td>0.070</td>
<td>0.285</td>
<td>0.112</td>
<td>0.426</td>
<td>0.077</td>
<td>0.152</td>
<td>0.132</td>
<td>0.065</td>
<td>0.086</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.117</td>
<td>0.993</td>
<td>0.066</td>
<td>0.209</td>
<td>0.118</td>
<td>0.466</td>
<td>0.070</td>
<td>0.155</td>
<td>0.142</td>
<td>0.066</td>
<td>0.078</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.162</td>
<td>0.866</td>
<td>0.060</td>
<td>0.480</td>
<td>0.148</td>
<td>0.422</td>
<td>0.060</td>
<td>0.638</td>
<td>0.132</td>
<td>0.057</td>
<td>0.076</td>
<td>0.473</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.181</td>
<td>0.941</td>
<td>0.053</td>
<td>0.602</td>
<td>0.147</td>
<td>0.533</td>
<td>0.056</td>
<td>0.918</td>
<td>0.394</td>
<td>0.097</td>
<td>0.062</td>
<td>0.931</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.107</td>
<td>1.00</td>
<td>0.062</td>
<td>0.213</td>
<td>0.110</td>
<td>0.653</td>
<td>0.063</td>
<td>0.141</td>
<td>0.128</td>
<td>0.054</td>
<td>0.067</td>
<td>0.565</td>
</tr>
<tr>
<td>$T$</td>
<td>4</td>
<td>0.108</td>
<td>1.00</td>
<td>0.064</td>
<td>0.335</td>
<td>0.109</td>
<td>0.596</td>
<td>0.065</td>
<td>0.149</td>
<td>0.130</td>
<td>0.055</td>
<td>0.067</td>
<td>0.268</td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>AIC</td>
<td>0.114</td>
<td>1.00</td>
<td>0.062</td>
<td>0.219</td>
<td>0.118</td>
<td>0.651</td>
<td>0.061</td>
<td>0.144</td>
<td>0.136</td>
<td>0.061</td>
<td>0.068</td>
<td>0.546</td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.147</td>
<td>0.974</td>
<td>0.056</td>
<td>0.398</td>
<td>0.127</td>
<td>0.613</td>
<td>0.060</td>
<td>0.458</td>
<td>0.128</td>
<td>0.054</td>
<td>0.067</td>
<td>0.555</td>
<td></td>
</tr>
<tr>
<td>VAR(3)–(MA2)</td>
<td>2</td>
<td>0.448</td>
<td>0.932</td>
<td>0.197</td>
<td>0.978</td>
<td>0.464</td>
<td>0.149</td>
<td>0.114</td>
<td>0.626</td>
<td>0.865</td>
<td>0.483</td>
<td>0.227</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.242</td>
<td>0.977</td>
<td>0.143</td>
<td>0.516</td>
<td>0.221</td>
<td>0.457</td>
<td>0.105</td>
<td>0.411</td>
<td>0.431</td>
<td>0.110</td>
<td>0.154</td>
<td>0.799</td>
</tr>
<tr>
<td>VAR(3)–(MA2)</td>
<td>$T=100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.187</td>
<td>0.985</td>
<td>0.129</td>
<td>0.564</td>
<td>0.177</td>
<td>0.812</td>
<td>0.114</td>
<td>0.480</td>
<td>0.226</td>
<td>0.152</td>
<td>0.144</td>
<td>0.815</td>
</tr>
<tr>
<td>AIC</td>
<td>0.254</td>
<td>0.964</td>
<td>0.161</td>
<td>0.585</td>
<td>0.239</td>
<td>0.835</td>
<td>0.161</td>
<td>0.600</td>
<td>0.326</td>
<td>0.178</td>
<td>0.205</td>
<td>0.846</td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.295</td>
<td>0.971</td>
<td>0.142</td>
<td>0.613</td>
<td>0.221</td>
<td>0.570</td>
<td>0.109</td>
<td>0.441</td>
<td>0.437</td>
<td>0.123</td>
<td>0.151</td>
<td>0.810</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.603</td>
<td>0.998</td>
<td>0.266</td>
<td>0.989</td>
<td>0.517</td>
<td>0.194</td>
<td>0.097</td>
<td>0.738</td>
<td>0.946</td>
<td>0.738</td>
<td>0.259</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.283</td>
<td>1.00</td>
<td>0.146</td>
<td>0.598</td>
<td>0.229</td>
<td>0.733</td>
<td>0.082</td>
<td>0.477</td>
<td>0.553</td>
<td>0.131</td>
<td>0.194</td>
</tr>
<tr>
<td>VAR(3)–(MA2)</td>
<td>$T=200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.191</td>
<td>1.00</td>
<td>0.099</td>
<td>0.814</td>
<td>0.158</td>
<td>0.988</td>
<td>0.085</td>
<td>0.578</td>
<td>0.290</td>
<td>0.212</td>
<td>0.123</td>
<td>0.908</td>
</tr>
<tr>
<td>AIC</td>
<td>0.166</td>
<td>1.00</td>
<td>0.101</td>
<td>0.678</td>
<td>0.165</td>
<td>0.987</td>
<td>0.101</td>
<td>0.744</td>
<td>0.182</td>
<td>0.117</td>
<td>0.123</td>
<td>0.871</td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.251</td>
<td>1.00</td>
<td>0.115</td>
<td>0.693</td>
<td>0.165</td>
<td>0.988</td>
<td>0.086</td>
<td>0.589</td>
<td>0.461</td>
<td>0.194</td>
<td>0.139</td>
<td>0.921</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.718</td>
<td>1.00</td>
<td>0.352</td>
<td>1.00</td>
<td>0.539</td>
<td>0.244</td>
<td>0.095</td>
<td>0.792</td>
<td>0.969</td>
<td>0.858</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.369</td>
<td>1.00</td>
<td>0.159</td>
<td>0.671</td>
<td>0.230</td>
<td>0.900</td>
<td>0.080</td>
<td>0.535</td>
<td>0.613</td>
<td>0.166</td>
<td>0.237</td>
</tr>
<tr>
<td>VAR(3)–(MA2)</td>
<td>$T=300$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.200</td>
<td>1.00</td>
<td>0.090</td>
<td>0.953</td>
<td>0.155</td>
<td>1.00</td>
<td>0.073</td>
<td>0.672</td>
<td>0.372</td>
<td>0.302</td>
<td>0.139</td>
<td>0.946</td>
</tr>
<tr>
<td>AIC</td>
<td>0.136</td>
<td>1.00</td>
<td>0.075</td>
<td>0.714</td>
<td>0.149</td>
<td>0.998</td>
<td>0.092</td>
<td>0.835</td>
<td>0.169</td>
<td>0.110</td>
<td>0.117</td>
<td>0.887</td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.223</td>
<td>1.00</td>
<td>0.094</td>
<td>0.899</td>
<td>0.152</td>
<td>1.00</td>
<td>0.074</td>
<td>0.733</td>
<td>0.348</td>
<td>0.220</td>
<td>0.130</td>
<td>0.974</td>
<td></td>
</tr>
</tbody>
</table>

a The tests examined are: $\lambda_{ST}$ (Wald test based on levels OLS regression); $\lambda_{DS}$ (Wald test based on differences OLS regression); $\lambda_{LW}$ (Wald test based on levels OLS regression with surplus lags added to isolate nonstandard asymptotics); and $\lambda_{FM}$ (Wald test based on fully modified vector autoregression). For each test, 5% $\chi^2$ critical values are used. The lags used are listed, as is the number of observations in the sample, and the DGP used to generate the data. All results are based on 5000 Monte Carlo trials. Entries in the table correspond to the proportion of rejections of the null hypothesis of noncasuality. For example, for the (CI-1)—VAR(1)—(MA1) with $T=100$ observations, the upper left entry of the table is 0.167. Thus, a standard Wald test when applied to the data in levels has an empirical rejection rate of 16.7%. 
Table 2: Summary of Sequential Wald Test Monte Carlo Results, Empirical Size; a Data-Generating Process is (CI-1)

<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>( \lambda_{p1} ) (VAR(1) MA1)</th>
<th>( \lambda_{p1} ) (VAR(1) MA2)</th>
<th>( \lambda_{p2} ) (VAR(1) MA1)</th>
<th>( \lambda_{p2} ) (VAR(1) MA2)</th>
<th>( \lambda_{p3} ) (VAR(1) MA1)</th>
<th>( \lambda_{p3} ) (VAR(1) MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.067</td>
<td>0.069</td>
<td>0.887</td>
<td>0.948</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.162</td>
<td>0.237</td>
<td>0.230</td>
<td>0.328</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.068</td>
<td>0.067</td>
<td>0.889</td>
<td>0.946</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.095</td>
<td>0.124</td>
<td>0.063</td>
<td>0.338</td>
<td>0.112</td>
<td>0.151</td>
<td>0.094</td>
</tr>
<tr>
<td>5</td>
<td>0.150</td>
<td>0.208</td>
<td>0.181</td>
<td>0.448</td>
<td>0.155</td>
<td>0.227</td>
<td>0.191</td>
</tr>
<tr>
<td>6</td>
<td>0.095</td>
<td>0.123</td>
<td>0.061</td>
<td>0.330</td>
<td>0.112</td>
<td>0.150</td>
<td>0.093</td>
</tr>
<tr>
<td>7</td>
<td>0.094</td>
<td>0.123</td>
<td>0.060</td>
<td>0.328</td>
<td>0.111</td>
<td>0.150</td>
<td>0.091</td>
</tr>
<tr>
<td>8</td>
<td>0.102</td>
<td>0.213</td>
<td>0.118</td>
<td>0.218</td>
<td>0.191</td>
<td>0.268</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.140</td>
<td>0.242</td>
<td>0.147</td>
<td>0.251</td>
<td>0.235</td>
<td>0.324</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.102</td>
<td>0.212</td>
<td>0.118</td>
<td>0.217</td>
<td>0.191</td>
<td>0.265</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.102</td>
<td>0.212</td>
<td>0.117</td>
<td>0.216</td>
<td>0.190</td>
<td>0.265</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.118</td>
<td>0.145</td>
<td>0.146</td>
<td>0.155</td>
<td>0.217</td>
<td>0.201</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.152</td>
<td>0.190</td>
<td>0.157</td>
<td>0.198</td>
<td>0.253</td>
<td>0.285</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.119</td>
<td>0.144</td>
<td>0.148</td>
<td>0.154</td>
<td>0.222</td>
<td>0.200</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.117</td>
<td>0.143</td>
<td>0.145</td>
<td>0.153</td>
<td>0.217</td>
<td>0.199</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.082</td>
<td>0.144</td>
<td>0.114</td>
<td>0.201</td>
<td>0.628</td>
<td>0.174</td>
<td>0.137</td>
</tr>
<tr>
<td>17</td>
<td>0.174</td>
<td>0.215</td>
<td>0.168</td>
<td>0.265</td>
<td>0.241</td>
<td>0.225</td>
<td>0.175</td>
</tr>
<tr>
<td>18</td>
<td>0.082</td>
<td>0.141</td>
<td>0.113</td>
<td>0.194</td>
<td>0.631</td>
<td>0.170</td>
<td>0.134</td>
</tr>
<tr>
<td>19</td>
<td>0.082</td>
<td>0.141</td>
<td>0.112</td>
<td>0.195</td>
<td>0.636</td>
<td>0.170</td>
<td>0.134</td>
</tr>
<tr>
<td>Lag</td>
<td>2</td>
<td>T = 100</td>
<td>$\hat{h}_1$</td>
<td>$\hat{h}_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.067</td>
<td>.068</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.126</td>
<td>.125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.069</td>
<td>.066</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.243</td>
<td>.234</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.887</td>
<td>.889</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.197</td>
<td>.196</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.101</td>
<td>.098</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.259</td>
<td>.250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.054</td>
<td>.054</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.068</td>
<td>.066</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.940</td>
<td>.940</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.964</td>
<td>.961</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lag</th>
<th>2</th>
<th>T = 200</th>
<th>$\hat{h}_1$</th>
<th>$\hat{h}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>.069</td>
<td>.069</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.147</td>
<td>.146</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.047</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.489</td>
<td>.486</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.088</td>
<td>.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.196</td>
<td>.196</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.078</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.568</td>
<td>.566</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.127</td>
<td>.127</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.207</td>
<td>.207</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.103</td>
<td>.103</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.581</td>
<td>.577</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lag</th>
<th>2</th>
<th>T = 200</th>
<th>$\hat{h}_1$</th>
<th>$\hat{h}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>.075</td>
<td>.075</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.231</td>
<td>.230</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.088</td>
<td>.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.254</td>
<td>.253</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.122</td>
<td>.122</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.303</td>
<td>.303</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.203</td>
<td>.203</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.353</td>
<td>.353</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lag</th>
<th>4</th>
<th>T = 200</th>
<th>$\hat{h}_1$</th>
<th>$\hat{h}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>.087</td>
<td>.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.127</td>
<td>.127</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.094</td>
<td>.094</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.133</td>
<td>.133</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.154</td>
<td>.154</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.161</td>
<td>.161</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>( \lambda_{p1} ) (VAR(1) MA1)</th>
<th>( \lambda_{p1} ) (VAR(1) MA2)</th>
<th>( \lambda_{p2} ) (VAR(1) MA1)</th>
<th>( \lambda_{p2} ) (VAR(1) MA2)</th>
<th>( \lambda_{p3} ) (VAR(1) MA1)</th>
<th>( \lambda_{p3} ) (VAR(1) MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.062</td>
<td>0.103</td>
<td>0.083</td>
<td>0.120</td>
<td>0.649</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.155</td>
<td>0.172</td>
<td>0.122</td>
<td>0.168</td>
<td>0.220</td>
<td>0.181</td>
</tr>
<tr>
<td>( T = 200 )</td>
<td>( j_{01} )</td>
<td>0.062</td>
<td>0.102</td>
<td>0.082</td>
<td>0.119</td>
<td>0.650</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>( j_{02} )</td>
<td>0.062</td>
<td>0.102</td>
<td>0.082</td>
<td>0.118</td>
<td>0.650</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.054</td>
<td>0.133</td>
<td>0.059</td>
<td>0.188</td>
<td>0.940</td>
<td>0.174</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>2</td>
<td>0.150</td>
<td>0.243</td>
<td>0.162</td>
<td>0.246</td>
<td>0.214</td>
<td>0.260</td>
</tr>
<tr>
<td>( T = 200 )</td>
<td>( j_{01} )</td>
<td>0.054</td>
<td>0.132</td>
<td>0.058</td>
<td>0.187</td>
<td>0.940</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>( j_{02} )</td>
<td>0.054</td>
<td>0.132</td>
<td>0.058</td>
<td>0.186</td>
<td>0.940</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.057</td>
<td>0.049</td>
<td>0.953</td>
<td>0.957</td>
<td>0.953</td>
<td>0.957</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>2</td>
<td>0.160</td>
<td>0.245</td>
<td>0.225</td>
<td>0.344</td>
<td>0.225</td>
<td>0.344</td>
</tr>
<tr>
<td>( T = 300 )</td>
<td>( j_{01} )</td>
<td>0.057</td>
<td>0.047</td>
<td>0.953</td>
<td>0.955</td>
<td>0.953</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>( j_{02} )</td>
<td>0.057</td>
<td>0.047</td>
<td>0.953</td>
<td>0.954</td>
<td>0.953</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.061</td>
<td>0.197</td>
<td>0.041</td>
<td>0.617</td>
<td>0.079</td>
<td>0.263</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>2</td>
<td>0.123</td>
<td>0.311</td>
<td>0.178</td>
<td>0.716</td>
<td>0.130</td>
<td>0.342</td>
</tr>
<tr>
<td>( T = 300 )</td>
<td>( j_{01} )</td>
<td>0.061</td>
<td>0.196</td>
<td>0.041</td>
<td>0.614</td>
<td>0.079</td>
<td>0.263</td>
</tr>
<tr>
<td></td>
<td>( j_{02} )</td>
<td>0.061</td>
<td>0.196</td>
<td>0.041</td>
<td>0.617</td>
<td>0.079</td>
<td>0.263</td>
</tr>
</tbody>
</table>
Causality and Predictive Accuracy Tests

<table>
<thead>
<tr>
<th>Lag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.055 0.291</td>
<td>0.072 0.329</td>
<td>0.120 0.349</td>
<td>0.203 0.423</td>
</tr>
<tr>
<td>1</td>
<td>0.107 0.361</td>
<td>0.112 0.381</td>
<td>0.203 0.423</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.055 0.290</td>
<td>0.072 0.329</td>
<td>0.120 0.349</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.055 0.290</td>
<td>0.072 0.329</td>
<td>0.120 0.349</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.111 0.194</td>
<td>0.116 0.203</td>
<td>0.212 0.285</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.071 0.137</td>
<td>0.084 0.146</td>
<td>0.137 0.164</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.071 0.137</td>
<td>0.084 0.146</td>
<td>0.137 0.164</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.064 0.082</td>
<td>0.062 0.089</td>
<td>0.080 0.112</td>
<td>0.291 0.119</td>
</tr>
<tr>
<td>8</td>
<td>0.167 0.139</td>
<td>0.113 0.143</td>
<td>0.231 0.146</td>
<td>0.120 0.147</td>
</tr>
<tr>
<td>9</td>
<td>0.064 0.082</td>
<td>0.062 0.099</td>
<td>0.080 0.112</td>
<td>0.288 0.119</td>
</tr>
<tr>
<td>10</td>
<td>0.057 0.116</td>
<td>0.054 0.157</td>
<td>0.953 0.155</td>
<td>0.077 0.169</td>
</tr>
<tr>
<td>11</td>
<td>0.160 0.209</td>
<td>0.146 0.217</td>
<td>0.225 0.226</td>
<td>0.152 0.227</td>
</tr>
<tr>
<td>12</td>
<td>0.057 0.115</td>
<td>0.054 0.157</td>
<td>0.953 0.155</td>
<td>0.077 0.169</td>
</tr>
<tr>
<td>13</td>
<td>0.057 0.115</td>
<td>0.053 0.156</td>
<td>0.953 0.155</td>
<td>0.076 0.169</td>
</tr>
</tbody>
</table>

*See footnotes to Table 1. The tests examined are: $\lambda_{p1-3}$ (sequential Wald tests based on VEC models estimated using Johansen's maximum likelihood procedure).
### Table 3  Summary of Sequential Wald Test Monte Carlo Results, Empirical Size; a Data-Generating Process is (CI-2)

<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>VAR(1) (MA1)</th>
<th>VAR(1) (MA2)</th>
<th>VAR(1) (MA1)</th>
<th>VAR(1) (MA2)</th>
<th>VAR(3) (MA1)</th>
<th>VAR(3) (MA2)</th>
<th>VAR(3) (MA1)</th>
<th>VAR(3) (MA2)</th>
<th>VAR(3) (MA1)</th>
<th>VAR(3) (MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.061</td>
<td>0.013</td>
<td>—</td>
<td>—</td>
<td>0.908</td>
<td>0.970</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.147</td>
<td>0.040</td>
<td>—</td>
<td>—</td>
<td>0.213</td>
<td>0.530</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.061</td>
<td>0.013</td>
<td>—</td>
<td>—</td>
<td>0.908</td>
<td>0.969</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.147</td>
<td>0.040</td>
<td>—</td>
<td>—</td>
<td>0.213</td>
<td>0.530</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.065</td>
<td>0.037</td>
<td>0.069</td>
<td>0.105</td>
<td>0.083</td>
<td>0.069</td>
<td>0.089</td>
<td>0.162</td>
<td>0.140</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.137</td>
<td>0.211</td>
<td>0.163</td>
<td>0.470</td>
<td>0.141</td>
<td>0.221</td>
<td>0.174</td>
<td>0.480</td>
<td>0.206</td>
<td>0.295</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>0.065</td>
<td>0.037</td>
<td>0.058</td>
<td>0.087</td>
<td>0.082</td>
<td>0.069</td>
<td>0.088</td>
<td>0.148</td>
<td>0.139</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.137</td>
<td>0.211</td>
<td>0.163</td>
<td>0.470</td>
<td>0.141</td>
<td>0.221</td>
<td>0.174</td>
<td>0.480</td>
<td>0.206</td>
<td>0.295</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>0.071</td>
<td>0.115</td>
<td>0.093</td>
<td>0.163</td>
<td>0.603</td>
<td>0.151</td>
<td>0.119</td>
<td>0.203</td>
<td>0.316</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.157</td>
<td>0.210</td>
<td>0.166</td>
<td>0.243</td>
<td>0.221</td>
<td>0.216</td>
<td>0.175</td>
<td>0.251</td>
<td>0.395</td>
<td>0.296</td>
</tr>
<tr>
<td>Lag</td>
<td>SIC</td>
<td>T</td>
<td>( j_01 )</td>
<td>( j_02 )</td>
<td>( j_1 )</td>
<td>( j_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>---</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.061</td>
<td>0.048</td>
<td>0.061</td>
<td>0.085</td>
<td>0.908</td>
<td>0.078</td>
<td>0.094</td>
<td>0.110</td>
<td>--</td>
<td>0.086</td>
<td>0.127</td>
</tr>
<tr>
<td>2</td>
<td>0.147</td>
<td>0.208</td>
<td>0.172</td>
<td>0.221</td>
<td>0.213</td>
<td>0.215</td>
<td>0.183</td>
<td>0.228</td>
<td>--</td>
<td>0.286</td>
<td>0.253</td>
</tr>
</tbody>
</table>

\( \hat{T} = 100 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.061</td>
<td>0.047</td>
<td>0.061</td>
<td>0.081</td>
<td>0.910</td>
<td>0.077</td>
</tr>
<tr>
<td>2</td>
<td>0.157</td>
<td>0.428</td>
<td>--</td>
<td>--</td>
<td>0.225</td>
<td>0.562</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.066</td>
<td>0.008</td>
<td>--</td>
<td>--</td>
<td>0.917</td>
<td>0.953</td>
</tr>
<tr>
<td>2</td>
<td>0.156</td>
<td>0.008</td>
<td>--</td>
<td>--</td>
<td>0.917</td>
<td>0.952</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.061</td>
<td>0.025</td>
<td>0.045</td>
<td>0.111</td>
<td>0.082</td>
<td>0.054</td>
</tr>
<tr>
<td>2</td>
<td>0.129</td>
<td>0.209</td>
<td>0.157</td>
<td>0.525</td>
<td>0.137</td>
<td>0.218</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td>--</td>
<td>0.054</td>
<td>0.048</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>--</td>
<td>--</td>
<td>0.112</td>
<td>0.228</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td>--</td>
<td>0.063</td>
<td>0.053</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>--</td>
<td>--</td>
<td>0.116</td>
<td>0.155</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td>--</td>
<td>0.064</td>
<td>0.052</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>--</td>
<td>--</td>
<td>0.116</td>
<td>0.167</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.062</td>
<td>0.084</td>
<td>0.061</td>
<td>0.098</td>
<td>0.610</td>
<td>0.117</td>
</tr>
<tr>
<td>2</td>
<td>0.163</td>
<td>0.160</td>
<td>0.118</td>
<td>0.167</td>
<td>0.229</td>
<td>0.166</td>
</tr>
</tbody>
</table>

\( \hat{T} = 200 \)

<table>
<thead>
<tr>
<th>Lag</th>
<th>SIC</th>
<th>T</th>
<th>( j_01 )</th>
<th>( j_02 )</th>
<th>( j_1 )</th>
<th>( j_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.063</td>
<td>0.084</td>
<td>0.061</td>
<td>0.098</td>
<td>0.612</td>
<td>0.117</td>
</tr>
<tr>
<td>2</td>
<td>0.063</td>
<td>0.084</td>
<td>0.061</td>
<td>0.097</td>
<td>0.611</td>
<td>0.117</td>
</tr>
<tr>
<td>Lag/sample</td>
<td>CI rank</td>
<td>$\lambda_{p1}$ (VAR(1) MA1)</td>
<td>$\lambda_{p1}$ (VAR(1) MA2)</td>
<td>$\lambda_{p1}$ (VAR(3) MA1)</td>
<td>$\lambda_{p1}$ (VAR(3) MA2)</td>
<td>$\lambda_{p2}$ (VAR(1) MA1)</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.056</td>
<td>0.042</td>
<td>0.052</td>
<td>0.059</td>
<td>0.917</td>
</tr>
<tr>
<td>$T = 200$</td>
<td>2</td>
<td>0.157</td>
<td>0.172</td>
<td>0.151</td>
<td>0.167</td>
<td>0.225</td>
</tr>
<tr>
<td></td>
<td>/o1</td>
<td>0.056</td>
<td>0.043</td>
<td>0.052</td>
<td>0.058</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>/o2</td>
<td>0.056</td>
<td>0.042</td>
<td>0.051</td>
<td>0.057</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.048</td>
<td>0.006</td>
<td>-</td>
<td>-</td>
<td>0.938</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>2</td>
<td>0.147</td>
<td>0.447</td>
<td>-</td>
<td>-</td>
<td>0.209</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>/o1</td>
<td>0.049</td>
<td>0.006</td>
<td>-</td>
<td>-</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>/o2</td>
<td>0.049</td>
<td>0.006</td>
<td>-</td>
<td>-</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.056</td>
<td>0.024</td>
<td>0.042</td>
<td>0.134</td>
<td>0.075</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>2</td>
<td>0.122</td>
<td>0.220</td>
<td>0.156</td>
<td>0.549</td>
<td>0.130</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>/o1</td>
<td>0.055</td>
<td>0.024</td>
<td>0.042</td>
<td>0.130</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>/o2</td>
<td>0.055</td>
<td>0.023</td>
<td>0.042</td>
<td>0.131</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>0.058</td>
<td>0.046</td>
<td>-</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.113</td>
<td>0.230</td>
<td>-</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>/o1</td>
<td>-</td>
<td>-</td>
<td>0.058</td>
<td>0.045</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>/o2</td>
<td>-</td>
<td>-</td>
<td>0.058</td>
<td>0.046</td>
<td>-</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>Lag = AIC</td>
<td>Lag = SIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T = 300$</td>
<td>$T = 300$</td>
<td>$T = 300$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
<td>$\lambda_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.147</td>
<td>0.148</td>
<td>0.147</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.147</td>
<td>0.148</td>
<td>0.147</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.938</td>
<td>0.938</td>
<td>0.938</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.938</td>
<td>0.938</td>
<td>0.938</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.068</td>
<td>0.068</td>
<td>0.068</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.068</td>
<td>0.068</td>
<td>0.068</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.068</td>
<td>0.068</td>
<td>0.068</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See footnotes to Table 2.*
<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank (MA1)</th>
<th>VAR(1) (MA1)</th>
<th>VAR(1) (MA2)</th>
<th>VAR(3) (MA1)</th>
<th>VAR(3) (MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.498 0.960</td>
<td>0.246 0.856</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>0.846 0.997</td>
<td>0.688 0.951</td>
<td>0.537 0.947</td>
<td>0.491 0.935</td>
<td>0.491 0.935</td>
</tr>
<tr>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>0.452 0.799</td>
<td>0.195 0.724</td>
<td>0.337 0.947</td>
<td>0.491 0.935</td>
<td>0.491 0.935</td>
</tr>
<tr>
<td>2</td>
<td>0.759 0.985</td>
<td>0.888 0.969</td>
<td>0.888 0.969</td>
<td>0.888 0.969</td>
<td>0.888 0.969</td>
</tr>
<tr>
<td>3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>0.165 0.432</td>
<td>0.162 0.552</td>
<td>0.162 0.552</td>
<td>0.162 0.552</td>
<td>0.162 0.552</td>
</tr>
<tr>
<td>3</td>
<td>0.157 0.465</td>
<td>0.150 0.432</td>
<td>0.150 0.432</td>
<td>0.150 0.432</td>
<td>0.150 0.432</td>
</tr>
<tr>
<td>4</td>
<td>0.491 0.935</td>
<td>0.491 0.935</td>
<td>0.491 0.935</td>
<td>0.491 0.935</td>
<td>0.491 0.935</td>
</tr>
<tr>
<td>4</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
</tr>
<tr>
<td>4</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
</tr>
<tr>
<td>4</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
</tr>
<tr>
<td>4</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
</tr>
<tr>
<td>4</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
</tr>
<tr>
<td>4</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
</tr>
<tr>
<td>4</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
<td>0.362 0.870</td>
</tr>
<tr>
<td>4</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
<td>0.489 0.917</td>
</tr>
<tr>
<td>4</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
<td>0.917 0.995</td>
</tr>
</tbody>
</table>

**Table 4** Summary of Sequential Wald Test Monte Carlo Results, Empirical Size**a** Data-Generating Process is (DS1)
<table>
<thead>
<tr>
<th>Lag</th>
<th>T</th>
<th>j01</th>
<th>j02</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.498</td>
<td>0.463</td>
<td>0.481</td>
</tr>
<tr>
<td>2</td>
<td>0.182</td>
<td>0.557</td>
<td>0.663</td>
</tr>
<tr>
<td>3</td>
<td>0.164</td>
<td>0.585</td>
<td>0.687</td>
</tr>
<tr>
<td>4</td>
<td>0.086</td>
<td>0.750</td>
<td>0.750</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.393</td>
<td>0.415</td>
<td>0.406</td>
</tr>
<tr>
<td>7</td>
<td>0.198</td>
<td>0.213</td>
<td>0.206</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>$\lambda_{p1}$ (VAR(1) (MA1))</th>
<th>$\lambda_{p1}$ (VAR(1) (MA2))</th>
<th>$\lambda_{p1}$ (VAR(3) (MA1))</th>
<th>$\lambda_{p1}$ (VAR(3) (MA2))</th>
<th>$\lambda_{p2}$ (VAR(1) (MA1))</th>
<th>$\lambda_{p2}$ (VAR(1) (MA2))</th>
<th>$\lambda_{p2}$ (VAR(3) (MA1))</th>
<th>$\lambda_{p2}$ (VAR(3) (MA2))</th>
<th>$\lambda_{p3}$ (VAR(1) (MA1))</th>
<th>$\lambda_{p3}$ (VAR(1) (MA2))</th>
<th>$\lambda_{p3}$ (VAR(3) (MA1))</th>
<th>$\lambda_{p3}$ (VAR(3) (MA2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.490</td>
<td>0.491</td>
<td>0.147</td>
<td>0.524</td>
<td>0.267</td>
<td>0.473</td>
<td>0.144</td>
<td>0.525</td>
<td>-</td>
<td>0.548</td>
<td>0.299</td>
<td>0.619</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>2</td>
<td>0.451</td>
<td>0.432</td>
<td>0.136</td>
<td>0.456</td>
<td>0.203</td>
<td>0.419</td>
<td>0.140</td>
<td>0.463</td>
<td>-</td>
<td>0.449</td>
<td>0.205</td>
<td>0.515</td>
</tr>
<tr>
<td>$T = 200$</td>
<td>$j_0$</td>
<td>1.00</td>
<td>0.696</td>
<td>0.200</td>
<td>0.784</td>
<td>0.839</td>
<td>0.680</td>
<td>0.200</td>
<td>0.793</td>
<td>-</td>
<td>0.701</td>
<td>0.500</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>$j_1$</td>
<td>0.818</td>
<td>0.678</td>
<td>0.387</td>
<td>0.751</td>
<td>0.500</td>
<td>0.658</td>
<td>0.355</td>
<td>0.758</td>
<td>-</td>
<td>0.688</td>
<td>0.548</td>
<td>0.825</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.469</td>
<td>0.971</td>
<td>-</td>
<td>-</td>
<td>0.248</td>
<td>0.912</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>2</td>
<td>0.458</td>
<td>0.886</td>
<td>-</td>
<td>-</td>
<td>0.202</td>
<td>0.834</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>$j_0$</td>
<td>0.600</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>0.600</td>
<td>0.985</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$j_1$</td>
<td>0.696</td>
<td>0.997</td>
<td>-</td>
<td>-</td>
<td>0.533</td>
<td>0.979</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.151</td>
<td>0.757</td>
<td>0.514</td>
<td>0.985</td>
<td>0.148</td>
<td>0.729</td>
<td>0.508</td>
<td>0.985</td>
<td>0.270</td>
<td>0.777</td>
<td>0.623</td>
<td>0.988</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>2</td>
<td>0.145</td>
<td>0.696</td>
<td>0.408</td>
<td>0.970</td>
<td>0.146</td>
<td>0.664</td>
<td>0.407</td>
<td>0.970</td>
<td>0.185</td>
<td>0.716</td>
<td>0.464</td>
<td>0.977</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>$j_0$</td>
<td>0.667</td>
<td>0.776</td>
<td>0.732</td>
<td>0.987</td>
<td>0.667</td>
<td>0.760</td>
<td>0.727</td>
<td>0.987</td>
<td>0.833</td>
<td>0.790</td>
<td>0.790</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td>$j_1$</td>
<td>0.364</td>
<td>0.776</td>
<td>0.689</td>
<td>0.987</td>
<td>0.333</td>
<td>0.756</td>
<td>0.695</td>
<td>0.987</td>
<td>0.576</td>
<td>0.790</td>
<td>0.765</td>
<td>0.989</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>-</td>
<td>-</td>
<td>0.144</td>
<td>0.716</td>
<td>-</td>
<td>-</td>
<td>0.146</td>
<td>0.722</td>
<td>-</td>
<td>-</td>
<td>0.284</td>
<td>0.802</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.135</td>
<td>0.613</td>
<td>-</td>
<td>-</td>
<td>0.138</td>
<td>0.620</td>
<td>-</td>
<td>-</td>
<td>0.190</td>
<td>0.675</td>
</tr>
<tr>
<td>$T = 300$</td>
<td>$j_0$</td>
<td>-</td>
<td>-</td>
<td>0.333</td>
<td>0.803</td>
<td>-</td>
<td>-</td>
<td>0.333</td>
<td>0.808</td>
<td>-</td>
<td>-</td>
<td>0.444</td>
<td>0.872</td>
</tr>
<tr>
<td></td>
<td>$j_1$</td>
<td>-</td>
<td>-</td>
<td>0.296</td>
<td>0.783</td>
<td>-</td>
<td>-</td>
<td>0.296</td>
<td>0.789</td>
<td>-</td>
<td>-</td>
<td>0.370</td>
<td>0.858</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>0.133</td>
<td>0.447</td>
<td>0.134</td>
<td>0.428</td>
<td>0.290</td>
<td>0.531</td>
<td>0.125</td>
<td>0.381</td>
<td>0.127</td>
<td>0.377</td>
<td>0.198</td>
<td>0.429</td>
</tr>
<tr>
<td>---------</td>
<td>---</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>7 = 300</td>
<td>(j_{01})</td>
<td>0.417</td>
<td>0.536</td>
<td>0.417</td>
<td>0.497</td>
<td>0.417</td>
<td>0.610</td>
<td>0.365</td>
<td>0.519</td>
<td>0.387</td>
<td>0.485</td>
<td>0.452</td>
<td>0.592</td>
</tr>
<tr>
<td>(j_{02})</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td></td>
</tr>
<tr>
<td>(j_{02})</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td>0.571</td>
<td>0.800</td>
<td></td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>2</td>
<td>0.447</td>
<td>0.166</td>
<td>0.142</td>
<td>0.180</td>
<td>0.209</td>
<td>0.165</td>
<td>0.177</td>
<td>0.153</td>
<td>0.198</td>
<td>0.482</td>
<td>0.338</td>
<td>0.289</td>
</tr>
<tr>
<td>7 = 300</td>
<td>(j_{01})</td>
<td>0.571</td>
<td>0.371</td>
<td>0.333</td>
<td>0.328</td>
<td>0.667</td>
<td>0.371</td>
<td>0.333</td>
<td>0.310</td>
<td>0.667</td>
<td>0.427</td>
<td>0.444</td>
<td>0.474</td>
</tr>
<tr>
<td>(j_{02})</td>
<td>0.654</td>
<td>0.382</td>
<td>0.310</td>
<td>0.306</td>
<td>0.387</td>
<td>0.310</td>
<td>0.301</td>
<td>0.500</td>
<td>0.462</td>
<td>0.414</td>
<td>0.451</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(j_{02})</td>
<td>0.469</td>
<td>0.374</td>
<td>0.144</td>
<td>0.404</td>
<td>0.286</td>
<td>0.146</td>
<td>0.390</td>
<td>0.459</td>
<td>0.284</td>
<td>0.505</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>2</td>
<td>0.458</td>
<td>0.336</td>
<td>0.135</td>
<td>0.357</td>
<td>0.202</td>
<td>0.329</td>
<td>0.138</td>
<td>0.266</td>
<td>0.373</td>
<td>0.190</td>
<td>0.408</td>
<td></td>
</tr>
<tr>
<td>7 = 300</td>
<td>(j_{01})</td>
<td>0.600</td>
<td>0.651</td>
<td>0.333</td>
<td>0.678</td>
<td>0.600</td>
<td>0.653</td>
<td>0.333</td>
<td>0.646</td>
<td>0.685</td>
<td>0.444</td>
<td>0.725</td>
<td></td>
</tr>
<tr>
<td>(j_{02})</td>
<td>0.696</td>
<td>0.593</td>
<td>0.296</td>
<td>0.641</td>
<td>0.533</td>
<td>0.587</td>
<td>0.296</td>
<td>0.613</td>
<td>0.637</td>
<td>0.370</td>
<td>0.691</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See footnotes to Table 2.*
<table>
<thead>
<tr>
<th>DGP/sample</th>
<th>Lag</th>
<th>AIC</th>
<th>SIC</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
<th>AIC</th>
<th>SIC</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR(1)–(MA1)</td>
<td>1</td>
<td>0.958</td>
<td>0.995</td>
<td>0.963</td>
<td>0.047</td>
<td>—</td>
<td>0.995</td>
<td>0.005</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.358</td>
<td>0.113</td>
<td>0.889</td>
<td>0.105</td>
<td>0.006</td>
<td>0.929</td>
<td>0.052</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 100</td>
<td>AIC</td>
<td>0.945</td>
<td>0.960</td>
<td>0.946</td>
<td>0.048</td>
<td>0.006</td>
<td>0.984</td>
<td>0.008</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.958</td>
<td>0.995</td>
<td>0.963</td>
<td>0.047</td>
<td>—</td>
<td>0.995</td>
<td>0.005</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.963</td>
<td>0.995</td>
<td>0.962</td>
<td>0.038</td>
<td>—</td>
<td>0.995</td>
<td>0.005</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 200</td>
<td>AIC</td>
<td>0.945</td>
<td>0.965</td>
<td>0.957</td>
<td>0.042</td>
<td>0.001</td>
<td>0.991</td>
<td>0.008</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.963</td>
<td>0.995</td>
<td>0.962</td>
<td>0.038</td>
<td>—</td>
<td>0.995</td>
<td>0.005</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.973</td>
<td>0.999</td>
<td>0.973</td>
<td>0.027</td>
<td>—</td>
<td>0.999</td>
<td>0.001</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 300</td>
<td>AIC</td>
<td>0.950</td>
<td>0.957</td>
<td>0.967</td>
<td>0.033</td>
<td>0.000</td>
<td>0.998</td>
<td>0.002</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.973</td>
<td>0.999</td>
<td>0.973</td>
<td>0.027</td>
<td>—</td>
<td>0.999</td>
<td>0.001</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR(1)–(MA2)</td>
<td>1</td>
<td>0.962</td>
<td>0.894</td>
<td>0.948</td>
<td>0.052</td>
<td>—</td>
<td>0.994</td>
<td>0.006</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.450</td>
<td>0.143</td>
<td>0.919</td>
<td>0.080</td>
<td>0.001</td>
<td>0.972</td>
<td>0.023</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 100</td>
<td>AIC</td>
<td>0.366</td>
<td>0.051</td>
<td>0.829</td>
<td>0.134</td>
<td>0.037</td>
<td>0.920</td>
<td>0.063</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.583</td>
<td>0.359</td>
<td>0.922</td>
<td>0.077</td>
<td>0.001</td>
<td>0.974</td>
<td>0.021</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.967</td>
<td>0.994</td>
<td>0.955</td>
<td>0.045</td>
<td>—</td>
<td>0.994</td>
<td>0.006</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 200</td>
<td>AIC</td>
<td>0.505</td>
<td>0.136</td>
<td>0.936</td>
<td>0.064</td>
<td>0.000</td>
<td>0.986</td>
<td>0.014</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.261</td>
<td>0.015</td>
<td>0.866</td>
<td>0.117</td>
<td>0.017</td>
<td>0.955</td>
<td>0.039</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.459</td>
<td>0.106</td>
<td>0.934</td>
<td>0.066</td>
<td>0.000</td>
<td>0.981</td>
<td>0.017</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 300</td>
<td>1</td>
<td>0.980</td>
<td>0.998</td>
<td>0.965</td>
<td>0.035</td>
<td>—</td>
<td>0.998</td>
<td>0.002</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.586</td>
<td>0.162</td>
<td>0.966</td>
<td>0.034</td>
<td>0.000</td>
<td>0.994</td>
<td>0.006</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.239</td>
<td>0.003</td>
<td>0.897</td>
<td>0.098</td>
<td>0.005</td>
<td>0.966</td>
<td>0.033</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.425</td>
<td>0.056</td>
<td>0.943</td>
<td>0.057</td>
<td>0.000</td>
<td>0.985</td>
<td>0.015</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR(3)-(MA1)</td>
<td>2</td>
<td>0.371</td>
<td>0.096</td>
<td>0.920</td>
<td>0.080</td>
<td>0.000</td>
<td>0.982</td>
<td>0.018</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 100</td>
<td>3</td>
<td>0.275</td>
<td>0.047</td>
<td>0.835</td>
<td>0.163</td>
<td>0.002</td>
<td>0.907</td>
<td>0.089</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.250</td>
<td>0.027</td>
<td>0.815</td>
<td>0.172</td>
<td>0.013</td>
<td>0.870</td>
<td>0.123</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.326</td>
<td>0.063</td>
<td>0.853</td>
<td>0.144</td>
<td>0.003</td>
<td>0.905</td>
<td>0.091</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.379</td>
<td>0.114</td>
<td>0.913</td>
<td>0.087</td>
<td>0.000</td>
<td>0.973</td>
<td>0.027</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.383</td>
<td>0.058</td>
<td>0.933</td>
<td>0.067</td>
<td>0.000</td>
<td>0.986</td>
<td>0.014</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 200</td>
<td>3</td>
<td>0.278</td>
<td>0.024</td>
<td>0.883</td>
<td>0.117</td>
<td>0.000</td>
<td>0.936</td>
<td>0.064</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.232</td>
<td>0.013</td>
<td>0.860</td>
<td>0.139</td>
<td>0.001</td>
<td>0.918</td>
<td>0.082</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.290</td>
<td>0.029</td>
<td>0.886</td>
<td>0.114</td>
<td>0.000</td>
<td>0.934</td>
<td>0.065</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.381</td>
<td>0.054</td>
<td>0.929</td>
<td>0.071</td>
<td>0.000</td>
<td>0.970</td>
<td>0.030</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.390</td>
<td>0.066</td>
<td>0.925</td>
<td>0.075</td>
<td>0.000</td>
<td>0.996</td>
<td>0.004</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 300</td>
<td>3</td>
<td>0.260</td>
<td>0.018</td>
<td>0.873</td>
<td>0.127</td>
<td>0.000</td>
<td>0.950</td>
<td>0.050</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.224</td>
<td>0.002</td>
<td>0.850</td>
<td>0.150</td>
<td>0.000</td>
<td>0.936</td>
<td>0.064</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.264</td>
<td>0.016</td>
<td>0.876</td>
<td>0.124</td>
<td>0.000</td>
<td>0.953</td>
<td>0.047</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.336</td>
<td>0.043</td>
<td>0.897</td>
<td>0.103</td>
<td>0.000</td>
<td>0.967</td>
<td>0.033</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR(3)-(MA2)</td>
<td>2</td>
<td>0.669</td>
<td>0.364</td>
<td>0.956</td>
<td>0.044</td>
<td>0.000</td>
<td>0.975</td>
<td>0.025</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T = 100</td>
<td>3</td>
<td>0.433</td>
<td>0.091</td>
<td>0.915</td>
<td>0.082</td>
<td>0.003</td>
<td>0.927</td>
<td>0.062</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.304</td>
<td>0.032</td>
<td>0.854</td>
<td>0.119</td>
<td>0.027</td>
<td>0.925</td>
<td>0.056</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>0.323</td>
<td>0.022</td>
<td>0.814</td>
<td>0.150</td>
<td>0.036</td>
<td>0.910</td>
<td>0.080</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIC</td>
<td>0.504</td>
<td>0.147</td>
<td>0.925</td>
<td>0.068</td>
<td>0.007</td>
<td>0.943</td>
<td>0.043</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Table 5 (Continued)

<table>
<thead>
<tr>
<th>DGP/sample</th>
<th>Lag</th>
<th>AIC</th>
<th>SIC</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
<th>Model Selected—AIC</th>
<th>Model Selected—SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>0.828</td>
<td>0.483</td>
<td>0.960</td>
<td>0.040</td>
<td>0.000</td>
<td>0.988</td>
<td>0.012</td>
</tr>
<tr>
<td>T = 200</td>
<td>3</td>
<td>0.485</td>
<td>0.062</td>
<td>0.936</td>
<td>0.064</td>
<td>0.000</td>
<td>0.960</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.312</td>
<td>0.006</td>
<td>0.915</td>
<td>0.085</td>
<td>0.000</td>
<td>0.964</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.225</td>
<td>0.001</td>
<td>0.854</td>
<td>0.134</td>
<td>0.012</td>
<td>0.956</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.412</td>
<td>0.029</td>
<td>0.936</td>
<td>0.064</td>
<td>0.000</td>
<td>0.960</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.895</td>
<td>0.385</td>
<td>0.976</td>
<td>0.024</td>
<td>0.000</td>
<td>0.992</td>
<td>0.008</td>
</tr>
<tr>
<td>T = 300</td>
<td>3</td>
<td>0.562</td>
<td>0.082</td>
<td>0.948</td>
<td>0.052</td>
<td>0.000</td>
<td>0.968</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.312</td>
<td>0.001</td>
<td>0.922</td>
<td>0.078</td>
<td>0.000</td>
<td>0.971</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.185</td>
<td>0.000</td>
<td>0.860</td>
<td>0.131</td>
<td>0.009</td>
<td>0.952</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.350</td>
<td>0.011</td>
<td>0.928</td>
<td>0.072</td>
<td>0.000</td>
<td>0.975</td>
<td>0.025</td>
</tr>
</tbody>
</table>

*Results based on the model selection criteria-type approaches discussed above are reported in this table. The lags used are listed, as is the number of observations in the sample, and the DGP used to generate the data. The proportion of "wins" accruing to each type of estimated model (VEC model (CI)), stationary levels model (levels), and difference stationary model (difference) are listed in columns 5–10.
<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>VAR(1)-(MA1) AIC</th>
<th>SIC</th>
<th>VAR(1)-(MA2) AIC</th>
<th>SIC</th>
<th>VAR(3)-(MA1) AIC</th>
<th>SIC</th>
<th>VAR(3)-(MA2) AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.076</td>
<td>0.076</td>
<td>0.023</td>
<td>0.023</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.241</td>
<td>0.115</td>
<td>0.289</td>
<td>0.121</td>
<td>0.203</td>
<td>0.072</td>
<td>0.541</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.233</td>
<td>0.096</td>
<td>0.337</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.000</td>
<td>0.974</td>
<td>0.999</td>
<td>0.994</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.231</td>
<td>0.097</td>
<td>0.205</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.999</td>
<td>0.724</td>
<td>0.998</td>
<td>0.914</td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.104</td>
<td>0.091</td>
<td>0.260</td>
<td>0.059</td>
<td>0.252</td>
<td>0.108</td>
<td>0.240</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.926</td>
<td>0.999</td>
<td>0.954</td>
<td>0.991</td>
<td>0.675</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.076</td>
<td>0.076</td>
<td>0.248</td>
<td>0.113</td>
<td>0.206</td>
<td>0.078</td>
<td>0.388</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.997</td>
<td>0.998</td>
<td>0.991</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.061</td>
<td>0.061</td>
<td>0.023</td>
<td>0.023</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.228</td>
<td>0.086</td>
<td>0.372</td>
<td>0.129</td>
<td>0.212</td>
<td>0.047</td>
<td>0.729</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.212</td>
<td>0.070</td>
<td>0.403</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.189</td>
<td>0.075</td>
<td>0.220</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.000</td>
<td>0.996</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.084</td>
<td>0.067</td>
<td>0.177</td>
<td>0.026</td>
<td>0.223</td>
<td>0.074</td>
<td>0.154</td>
<td>0.024</td>
</tr>
</tbody>
</table>

(continued)
Table 6 (Continued)

<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>AIC (VAR(1) − (MA1))</th>
<th>SIC (VAR(1) − (MA1))</th>
<th>AIC (VAR(1) − (MA2))</th>
<th>SIC (VAR(1) − (MA2))</th>
<th>AIC (VAR(3) − (MA1))</th>
<th>SIC (VAR(3) − (MA1))</th>
<th>AIC (VAR(3) − (MA2))</th>
<th>SIC (VAR(3) − (MA2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>T = 200</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.982</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.887</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.061</td>
<td>0.061</td>
<td>0.339</td>
<td>0.101</td>
<td>0.233</td>
<td>0.059</td>
<td>0.322</td>
<td>0.048</td>
</tr>
<tr>
<td>T = 200</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.058</td>
<td>0.058</td>
<td>0.023</td>
<td>0.023</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.201</td>
<td>0.088</td>
<td>0.456</td>
<td>0.155</td>
<td>0.216</td>
<td>0.054</td>
<td>0.841</td>
<td>0.556</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.182</td>
<td>0.057</td>
<td>0.488</td>
<td>0.100</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.181</td>
<td>0.054</td>
<td>0.223</td>
<td>0.015</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.079</td>
<td>0.064</td>
<td>0.157</td>
<td>0.020</td>
<td>0.188</td>
<td>0.052</td>
<td>0.135</td>
<td>0.023</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.996</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.942</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.058</td>
<td>0.058</td>
<td>0.303</td>
<td>0.060</td>
<td>0.213</td>
<td>0.058</td>
<td>0.256</td>
<td>0.021</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*aSee Footnotes to Table 5. The reported numerical values are the proportion of rejections of the "null hypothesis" of noncausality. All models are compared in pairwise fashion. For example, in this table each entry corresponds to the rejection frequency when two alternative cointegrating models are compared—one with the causal variable of interest included, and one without.

*bAll estimations are based on the assumption that it is known that the data are generated according to a VEC.
models are estimated in levels, the tabulated size of the \textit{ex ante} tests often approaches zero very quickly as \( T \) increases, regardless of the true DGP.

7. Difference stationary models perform exceptionally poorly, in all cointegrated cases, suggesting that difference-type Wald tests do not provide a short cut to testing for “short-run” noncausality when the true DGP is a VEC. This is because least squares becomes very confused in the context of this type of mis-specification, and parameter estimates are hence very inaccurate. A final result worth noting is that even though the empirical sizes of many tests (with the exception of SIC and \textit{ex ante} type model selection based tests) are often too high, suggesting that the null of Granger causality may be rejected too often, the use of \( F \) versions of the Wald-type tests does not appear to decrease the empirical sizes to any great extent, for the sample sizes that we consider. For this reason, we report rejection frequencies based only on the \( \chi^2 \) versions of our test statistics.

\begin{table}[h]
\centering
\caption{SIC and AIC Approach Rejection Frequencies; Data-Generating Process is (CI-1)\textsuperscript{a}\textsuperscript{b}}
\begin{tabular}{cccccccc}
\hline
 & \multicolumn{2}{c}{VAR(1)--(MA1)} & \multicolumn{2}{c}{VAR(1)--(MA2)} & \multicolumn{2}{c}{VAR(3)--(MA1)} & \multicolumn{2}{c}{VAR(3)--(MA2)} \\
Sample & Lag & AIC & SIC & AIC & SIC & AIC & SIC & AIC & SIC \\
\hline
 & 1 & 0.329 & 0.113 & 0.410 & 0.205 & - & - & - & - \\
 & \( T = 100 \) & 2 & 0.262 & 0.036 & 0.386 & 0.082 & 0.321 & 0.005 & 0.596 & 0.248 \\
 & 3 & - & - & - & - & 0.212 & 0.010 & 0.370 & 0.037 \\
 & 4 & - & - & - & - & 0.202 & 0.006 & 0.261 & 0.018 \\
 & AIC & 0.341 & 0.113 & 0.306 & 0.020 & 0.254 & 0.017 & 0.263 & 0.010 \\
 & SIC & 0.329 & 0.113 & 0.410 & 0.114 & 2.326 & 0.058 & 0.417 & 0.092 \\
 & 1 & 0.359 & 0.081 & 0.420 & 0.156 & - & - & - & - \\
 & \( T = 200 \) & 2 & 0.272 & 0.018 & 0.427 & 0.063 & 0.315 & 0.033 & 0.759 & 0.300 \\
 & 3 & - & - & - & - & 0.205 & 0.012 & 0.432 & 0.034 \\
 & 4 & - & - & - & - & 0.177 & 0.003 & 0.256 & 0.003 \\
 & AIC & 0.363 & 0.081 & 0.212 & 0.005 & 0.214 & 0.011 & 0.178 & 0.002 \\
 & SIC & 0.359 & 0.081 & 0.399 & 0.048 & 0.307 & 0.030 & 0.358 & 0.019 \\
 & 1 & 0.349 & 0.089 & 0.458 & 0.168 & - & - & - & - \\
 & \( T = 300 \) & 2 & 0.264 & 0.014 & 0.506 & 0.079 & 0.350 & 0.031 & 0.825 & 0.368 \\
 & 3 & - & - & - & - & 0.193 & 0.003 & 0.496 & 0.040 \\
 & 4 & - & - & - & - & 0.166 & 0.000 & 0.268 & 0.00 \\
 & AIC & 0.354 & 0.089 & 0.198 & 0.001 & 0.199 & 0.003 & 0.154 & 0.00 \\
 & SIC & 0.349 & 0.089 & 0.352 & 0.028 & 0.276 & 0.022 & 0.299 & 0.004 \\
\hline
\end{tabular}
\end{table}

\textsuperscript{a}See footnotes to Table 7.

\textsuperscript{b}All estimations are based on the assumption that it is known that the data are generated according to a levels VAR.
4.1 Cointegration Models

In all of the cointegrated models (please see attached tables) the sequential Wald-type tests, the Toda and Yamamoto Wald-type tests, and the model selection based SIC type causality tests perform quite well. This is most evident when the true DGPs are VAR(1) and VAR(3) without an MA component. In these cases, \( \lambda_{LW} \) type tests have empirical sizes ranging from 0.049 to 0.056 for VAR(1)—\( T = 200 \) or 300, and approximately 0.055 for VAR(3)—\( T = 300 \) (see Table 1). In these cases, the best size is usually (except for VAR(1)–MA1) associated with lag selection based on the AIC. In fact, it turns out that the AIC is also useful for selecting lags in the context of our model selection based tests (see below). Thus, the more parsimonious method for choosing lags (SIC) performs better primarily for our least complex models. One reason for this finding may be that when complex DGPs are specified (as is usually the case when

<table>
<thead>
<tr>
<th>Sample</th>
<th>Lag</th>
<th>VAR(1)–(MA1) AIC</th>
<th>SIC</th>
<th>VAR(1)–(MA2) AIC</th>
<th>SIC</th>
<th>VAR(3)–(MA1) AIC</th>
<th>SIC</th>
<th>VAR(3)–(MA2) AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1.00</td>
<td>0.999</td>
<td>1.00</td>
<td>1.00</td>
<td>0.972</td>
<td>0.449</td>
<td>0.972</td>
<td>0.916</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.946</td>
<td>0.512</td>
<td>0.993</td>
<td>0.835</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1.00</td>
<td>0.945</td>
<td>0.998</td>
<td>0.922</td>
<td>0.892</td>
<td>0.609</td>
<td>0.977</td>
<td>0.631</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.737</td>
<td>0.467</td>
<td>0.986</td>
<td>0.897</td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>1.00</td>
<td>0.980</td>
<td>1.00</td>
<td>0.988</td>
<td>0.997</td>
<td>0.942</td>
<td>0.998</td>
<td>0.833</td>
</tr>
<tr>
<td>SIC</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.932</td>
<td>0.776</td>
<td>1.00</td>
<td>0.995</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.971</td>
<td>0.856</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.946</td>
<td>0.512</td>
<td>0.993</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1.00</td>
<td>0.998</td>
<td>1.00</td>
<td>1.00</td>
<td>0.993</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.987</td>
<td>0.938</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*a*See footnotes to Table 7.

*b*All estimations are based on the assumption that it is known that the data are generated according to a VAR in differences.
macroeconomic data are analysed), the specification of surplus lags helps standard maximum likelihood estimation techniques to become less confused. This point is particularly important, given the tendency of our simplest TYDL-type tests to perform so well. Put another way, if model complexity and estimation confusion is an issue, then avoiding the estimation of cointegrating vectors, ranks, etc., may be a useful approach—even if the true DGP is cointegrated. This indeed appears to be the case for more complex models, as the TYDL approach appears to perform as well as all of the other standard hypothesis testing approaches, at least in the context of empirical size. Of course, the cost of estimating models in levels, and ignoring cointegrating restrictions, is a reduction in efficiency.

Tests based on difference model estimations perform very poorly all the time. However, the Wald tests based on $\lambda_{ST}$ have sizes that are usually between 10 and 15%. Tests based on models estimated using differenced data perform very poorly whenever the true DGP is a VEC. However, $\lambda_{ST}$ tests have sizes that are usually between 10 and 15%, with the better sizes coming at the expense of larger samples, but not necessarily less model complexity. For example, in Table 1, for $T = 300$, the $\lambda_{ST}$ test has size = 0.108 in the VAR(3)—MA1—(CI-1) case when estimation is based on four lags, while the analogous test for the VAR(1) (i.e., two lags used in estimation) has size = 0.128. Note that the reason for the relatively good performance of these two models, each of which is estimated with an extra lag, is probably related to the argument used to show that the $\lambda_{LW}$ type tests are asymptotically valid. That is, even when we use the standard Wald test, as long as an extra lag is included, the size of the test is not too far from the nominal size, even though the asymptotic $\chi^2$ distribution is clearly not valid.

The results for the tests based on the $\lambda_{FM}$ statistic are somewhat less encouraging than those of the Toda and Yamamoto and sequential Wald-type tests. In all of our experiments, the size of these tests appears to be in the 15–30% range, regardless of sample size, model specification, etc. Although these results are not too bad, relatively, it seems apparent that a number of the other tests perform better. As the $\lambda_{FM}$ test statistics rely on some optimal choice of bandwidth parameter as well as kernel function, for example, our results should be viewed with caution, as more suitable FM–VAR procedures may be available. To try and avoid this problem somewhat, we used three different lag truncation parameters to define the number of autocovariance terms used to compute the spectrum at zero frequency. These have been used elsewhere, and are: (1) $l = 0$, (2) $l = \text{integer} \times (T/100)^{0.25}$, and (3) $l = \text{integer} \times (T/100)^{0.25}$. Further, we used the Parzen kernel function for the calculation of long-run covariance matrix estimates, as well as Andrews’ (1991) automatic bandwidth selection procedures. Finally, an AR(1) filter was used to flatten the spectrum of the error around the zero frequency. Our
### Table 9: MSE, MAD, and MAPE Type Ex Ante Predictive Ability: Data-Generating Process is (CI-1)$^b$

<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>VAR(1)-(MA1)</th>
<th></th>
<th>VAR(1)-(MA2)</th>
<th></th>
<th>VAR(3)-(MA1)</th>
<th></th>
<th>VAR(3)-(MA2)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>MAD</td>
<td>MAPE</td>
<td></td>
<td>MSE</td>
<td>MAD</td>
<td>MAPE</td>
<td></td>
</tr>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.166</td>
<td>0.158</td>
<td>0.114</td>
<td>0.288</td>
<td>0.243</td>
<td>0.185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 = 100</td>
<td>2</td>
<td>0.147</td>
<td>0.144</td>
<td>0.105</td>
<td>0.260</td>
<td>0.228</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.013</td>
<td>0.016</td>
<td>0.014</td>
<td>0.014</td>
<td>0.019</td>
<td>0.016</td>
<td>0.011</td>
<td>0.017</td>
</tr>
<tr>
<td>7 = 100</td>
<td>2</td>
<td>0.063</td>
<td>0.082</td>
<td>0.071</td>
<td>0.145</td>
<td>0.152</td>
<td>0.125</td>
<td>0.066</td>
<td>0.058</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7 = 100</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.008</td>
<td>0.010</td>
</tr>
<tr>
<td>7 = 100</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.164</td>
<td>0.153</td>
<td>0.111</td>
<td>0.005</td>
<td>0.007</td>
<td>0.007</td>
<td>0.010</td>
<td>0.017</td>
</tr>
<tr>
<td>7 = 100</td>
<td>2</td>
<td>0.144</td>
<td>0.142</td>
<td>0.105</td>
<td>0.080</td>
<td>0.071</td>
<td>0.065</td>
<td>0.048</td>
<td>0.051</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.166</td>
<td>0.158</td>
<td>0.114</td>
<td>0.077</td>
<td>0.077</td>
<td>0.068</td>
<td>0.014</td>
<td>0.018</td>
</tr>
<tr>
<td>7 = 100</td>
<td>2</td>
<td>0.147</td>
<td>0.144</td>
<td>0.105</td>
<td>0.173</td>
<td>0.180</td>
<td>0.138</td>
<td>0.070</td>
<td>0.061</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.595</td>
<td>0.519</td>
<td>0.373</td>
<td>0.716</td>
<td>0.630</td>
<td>0.441</td>
<td></td>
<td>—</td>
</tr>
<tr>
<td>7 = 200</td>
<td>2</td>
<td>0.564</td>
<td>0.495</td>
<td>0.357</td>
<td>0.713</td>
<td>0.625</td>
<td>0.431</td>
<td></td>
<td>—</td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.025</td>
<td>0.017</td>
<td>0.011</td>
<td>0.017</td>
<td>0.022</td>
<td>0.014</td>
<td>0.013</td>
<td>0.012</td>
</tr>
<tr>
<td>7 = 200</td>
<td>2</td>
<td>0.332</td>
<td>0.288</td>
<td>0.199</td>
<td>0.524</td>
<td>0.464</td>
<td>0.303</td>
<td>0.232</td>
<td>0.192</td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.008</td>
<td>0.004</td>
</tr>
<tr>
<td>7 = 200</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.224</td>
<td>0.171</td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.009</td>
<td>0.012</td>
</tr>
</tbody>
</table>
The approaches reported in this table are based on the use of ex ante MSE, MAD, and MAPE forecasting criteria. The lags used are listed, as is the number of observations in the sample. The DGPs used to generate the data are given across the columns of the table. Each entry corresponds to the rejection frequency with two alternative cointegrating models compared, one with the causal variable of interest included and another without. All models are compared in pairwise fashion. For example, in this table each entry corresponds to the rejection frequency when two alternative cointegrating models are compared, one with the causal variable of interest included and another without. All estimations are based on the assumption that it is known that the data are generated according to a VEC. All estimations are based on the assumption that it is known that the data are generated according to a VEC.
<table>
<thead>
<tr>
<th>Sample</th>
<th>MSE</th>
<th>MAD</th>
<th>MAPE</th>
<th>MSE</th>
<th>MAD</th>
<th>MAPE</th>
<th>MSE</th>
<th>MAD</th>
<th>MAPE</th>
<th>MSE</th>
<th>MAD</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>T = 100</td>
<td>0.018</td>
<td>0.012</td>
<td>0.024</td>
<td>0.016</td>
<td>0.018</td>
<td>0.021</td>
<td>0.013</td>
<td>0.015</td>
<td>0.013</td>
<td>0.021</td>
<td>0.021</td>
<td>0.013</td>
</tr>
<tr>
<td>2</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.016</td>
<td>0.016</td>
<td>0.014</td>
<td>0.015</td>
<td>0.012</td>
<td>0.016</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>3</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.016</td>
<td>0.016</td>
<td>0.014</td>
<td>0.015</td>
<td>0.012</td>
<td>0.016</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>4</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.016</td>
<td>0.016</td>
<td>0.014</td>
<td>0.015</td>
<td>0.012</td>
<td>0.016</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>AIC</td>
<td>0.018</td>
<td>0.014</td>
<td>0.025</td>
<td>0.003</td>
<td>0.005</td>
<td>0.009</td>
<td>0.015</td>
<td>0.013</td>
<td>0.014</td>
<td>0.008</td>
<td>0.011</td>
<td>0.013</td>
</tr>
<tr>
<td>SIC</td>
<td>0.018</td>
<td>0.012</td>
<td>0.024</td>
<td>0.019</td>
<td>0.021</td>
<td>0.017</td>
<td>0.013</td>
<td>0.016</td>
<td>0.014</td>
<td>0.010</td>
<td>0.013</td>
<td>0.009</td>
</tr>
<tr>
<td>T = 200</td>
<td>0.017</td>
<td>0.017</td>
<td>0.025</td>
<td>0.022</td>
<td>0.020</td>
<td>0.028</td>
<td>0.013</td>
<td>0.012</td>
<td>0.018</td>
<td>0.036</td>
<td>0.039</td>
<td>0.037</td>
</tr>
<tr>
<td>2</td>
<td>0.007</td>
<td>0.009</td>
<td>0.010</td>
<td>0.012</td>
<td>0.015</td>
<td>0.016</td>
<td>0.013</td>
<td>0.012</td>
<td>0.018</td>
<td>0.036</td>
<td>0.039</td>
<td>0.037</td>
</tr>
<tr>
<td>3</td>
<td>0.007</td>
<td>0.009</td>
<td>0.010</td>
<td>0.012</td>
<td>0.015</td>
<td>0.016</td>
<td>0.013</td>
<td>0.012</td>
<td>0.018</td>
<td>0.036</td>
<td>0.039</td>
<td>0.037</td>
</tr>
<tr>
<td>4</td>
<td>0.007</td>
<td>0.009</td>
<td>0.010</td>
<td>0.012</td>
<td>0.015</td>
<td>0.016</td>
<td>0.013</td>
<td>0.012</td>
<td>0.018</td>
<td>0.036</td>
<td>0.039</td>
<td>0.037</td>
</tr>
<tr>
<td>AIC</td>
<td>0.017</td>
<td>0.017</td>
<td>0.025</td>
<td>0.004</td>
<td>0.004</td>
<td>0.009</td>
<td>0.005</td>
<td>0.007</td>
<td>0.016</td>
<td>0.002</td>
<td>0.004</td>
<td>0.003</td>
</tr>
<tr>
<td>SIC</td>
<td>0.017</td>
<td>0.017</td>
<td>0.025</td>
<td>0.010</td>
<td>0.013</td>
<td>0.013</td>
<td>0.011</td>
<td>0.008</td>
<td>0.015</td>
<td>0.012</td>
<td>0.007</td>
<td>0.013</td>
</tr>
<tr>
<td>T = 300</td>
<td>0.027</td>
<td>0.020</td>
<td>0.023</td>
<td>0.027</td>
<td>0.031</td>
<td>0.039</td>
<td>0.010</td>
<td>0.012</td>
<td>0.018</td>
<td>0.047</td>
<td>0.049</td>
<td>0.026</td>
</tr>
<tr>
<td>2</td>
<td>0.009</td>
<td>0.013</td>
<td>0.013</td>
<td>0.010</td>
<td>0.014</td>
<td>0.014</td>
<td>0.006</td>
<td>0.011</td>
<td>0.017</td>
<td>0.013</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>3</td>
<td>0.009</td>
<td>0.013</td>
<td>0.013</td>
<td>0.010</td>
<td>0.014</td>
<td>0.014</td>
<td>0.006</td>
<td>0.011</td>
<td>0.017</td>
<td>0.013</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>4</td>
<td>0.009</td>
<td>0.013</td>
<td>0.013</td>
<td>0.010</td>
<td>0.014</td>
<td>0.014</td>
<td>0.006</td>
<td>0.011</td>
<td>0.017</td>
<td>0.013</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>AIC</td>
<td>0.026</td>
<td>0.020</td>
<td>0.022</td>
<td>0.002</td>
<td>0.012</td>
<td>0.008</td>
<td>0.005</td>
<td>0.011</td>
<td>0.017</td>
<td>0.003</td>
<td>0.004</td>
<td>0.010</td>
</tr>
<tr>
<td>SIC</td>
<td>0.027</td>
<td>0.020</td>
<td>0.023</td>
<td>0.006</td>
<td>0.008</td>
<td>0.012</td>
<td>0.013</td>
<td>0.012</td>
<td>0.018</td>
<td>0.003</td>
<td>0.009</td>
<td>0.013</td>
</tr>
</tbody>
</table>

*See footnotes to Table 9.*

*All estimations are based on the assumption that it is known that the data are generated according to a levels VAR in differences.*
<table>
<thead>
<tr>
<th>Sample</th>
<th>Lag</th>
<th>VAR(1)–(MA1)</th>
<th>VAR(1)–(MA2)</th>
<th>VAR(3)–(MA1)</th>
<th>VAR(3)–(MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>MAD</td>
<td>MAPE</td>
<td>MSE</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.068</td>
<td>0.064</td>
<td>0.041</td>
<td>0.083</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.043</td>
<td>0.039</td>
<td>0.039</td>
<td>0.083</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.038</td>
<td>0.033</td>
<td>0.032</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.055</td>
<td>0.00</td>
<td>0.00</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.180</td>
<td>0.175</td>
<td>0.128</td>
<td>0.301</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.081</td>
<td>0.077</td>
<td>0.047</td>
<td>0.142</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.088</td>
<td>0.072</td>
<td>0.060</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.180</td>
<td>0.120</td>
<td>0.110</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.371</td>
<td>0.298</td>
<td>0.190</td>
<td>0.542</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.162</td>
<td>0.141</td>
<td>0.092</td>
<td>0.257</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.159</td>
<td>0.119</td>
<td>0.088</td>
<td>0.259</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.412</td>
<td>0.353</td>
<td>0.157</td>
<td>0.349</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.505</td>
<td>0.390</td>
<td>0.280</td>
<td>0.131</td>
</tr>
</tbody>
</table>

*aSee footnotes to Table 9.

*bAll estimations are based on the assumption that it is known that the data are generated according to a VAR in differences.
<table>
<thead>
<tr>
<th>DGP/sample</th>
<th>Lag</th>
<th>AIC</th>
<th>SIC</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR(1)-(MA1)</td>
<td>2</td>
<td>0.388</td>
<td>0.123</td>
<td>0.858</td>
<td>0.117</td>
<td>0.025</td>
<td>0.911</td>
<td>0.044</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.931</td>
<td>0.950</td>
<td>0.953</td>
<td>0.043</td>
<td>0.004</td>
<td>0.983</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.962</td>
<td>0.994</td>
<td>0.961</td>
<td>0.039</td>
<td></td>
<td>0.994</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.951</td>
<td>0.994</td>
<td>0.949</td>
<td>0.051</td>
<td></td>
<td>0.994</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>VAR(1)-(MA2)</td>
<td>2</td>
<td>0.334</td>
<td>0.061</td>
<td>0.876</td>
<td>0.124</td>
<td>0.000</td>
<td>0.958</td>
<td>0.042</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.925</td>
<td>0.946</td>
<td>0.941</td>
<td>0.059</td>
<td>0.000</td>
<td>0.994</td>
<td>0.006</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.946</td>
<td>0.997</td>
<td>0.943</td>
<td>0.057</td>
<td></td>
<td>0.997</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>VAR(1)-(MA2)</td>
<td>1</td>
<td>0.946</td>
<td>0.984</td>
<td>0.913</td>
<td>0.087</td>
<td></td>
<td>0.980</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.353</td>
<td>0.046</td>
<td>0.804</td>
<td>0.134</td>
<td>0.062</td>
<td>0.919</td>
<td>0.056</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.436</td>
<td>0.121</td>
<td>0.892</td>
<td>0.101</td>
<td>0.007</td>
<td>0.971</td>
<td>0.019</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Causality and Predictive Accuracy Tests

<table>
<thead>
<tr>
<th>( T = 200 )</th>
<th>( T = 300 )</th>
<th>( T = 100 )</th>
<th>( \text{VAR}(3)-(MA1) )</th>
<th>( T = 200 )</th>
<th>( T = 300 )</th>
<th>( T = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.986</td>
<td>0.986</td>
<td>0.935</td>
<td>0.467</td>
<td>0.467</td>
<td>0.467</td>
</tr>
<tr>
<td>AIC</td>
<td>0.233</td>
<td>0.233</td>
<td>0.259</td>
<td>0.233</td>
<td>0.233</td>
<td>0.233</td>
</tr>
<tr>
<td>SIC</td>
<td>0.335</td>
<td>0.335</td>
<td>0.325</td>
<td>0.335</td>
<td>0.335</td>
<td>0.335</td>
</tr>
<tr>
<td>2</td>
<td>0.990</td>
<td>0.990</td>
<td>0.980</td>
<td>0.465</td>
<td>0.465</td>
<td>0.465</td>
</tr>
<tr>
<td>AIC</td>
<td>0.225</td>
<td>0.225</td>
<td>0.225</td>
<td>0.225</td>
<td>0.225</td>
<td>0.225</td>
</tr>
<tr>
<td>SIC</td>
<td>0.325</td>
<td>0.325</td>
<td>0.325</td>
<td>0.325</td>
<td>0.325</td>
<td>0.325</td>
</tr>
<tr>
<td>3</td>
<td>0.956</td>
<td>0.956</td>
<td>0.915</td>
<td>0.388</td>
<td>0.388</td>
<td>0.388</td>
</tr>
<tr>
<td>AIC</td>
<td>0.215</td>
<td>0.215</td>
<td>0.215</td>
<td>0.215</td>
<td>0.215</td>
<td>0.215</td>
</tr>
<tr>
<td>SIC</td>
<td>0.305</td>
<td>0.305</td>
<td>0.305</td>
<td>0.305</td>
<td>0.305</td>
<td>0.305</td>
</tr>
<tr>
<td>4</td>
<td>0.941</td>
<td>0.941</td>
<td>0.901</td>
<td>0.293</td>
<td>0.293</td>
<td>0.293</td>
</tr>
<tr>
<td>AIC</td>
<td>0.205</td>
<td>0.205</td>
<td>0.205</td>
<td>0.205</td>
<td>0.205</td>
<td>0.205</td>
</tr>
<tr>
<td>SIC</td>
<td>0.295</td>
<td>0.295</td>
<td>0.295</td>
<td>0.295</td>
<td>0.295</td>
<td>0.295</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>DGP / sample</th>
<th>Lag</th>
<th>AIC</th>
<th>SIC</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
<th>CI</th>
<th>Levels</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR(3)-(MA2)</td>
<td>2</td>
<td>0.671</td>
<td>0.381</td>
<td>0.912</td>
<td>0.088</td>
<td>0.000</td>
<td>0.932</td>
<td>0.068</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.404</td>
<td>0.049</td>
<td>0.927</td>
<td>0.072</td>
<td>0.001</td>
<td>0.978</td>
<td>0.021</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.276</td>
<td>0.017</td>
<td>0.908</td>
<td>0.087</td>
<td>0.005</td>
<td>0.975</td>
<td>0.019</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.936</td>
<td>0.056</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AIC</td>
<td>0.293</td>
<td>0.015</td>
<td>0.936</td>
<td>0.056</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SIC</td>
<td>0.376</td>
<td>0.051</td>
<td>0.936</td>
<td>0.056</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.376</td>
<td>0.051</td>
<td>0.009</td>
</tr>
<tr>
<td>T = 200</td>
<td>2</td>
<td>0.745</td>
<td>0.383</td>
<td>0.907</td>
<td>0.093</td>
<td>0.000</td>
<td>0.945</td>
<td>0.056</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.459</td>
<td>0.041</td>
<td>0.924</td>
<td>0.076</td>
<td>0.000</td>
<td>0.988</td>
<td>0.012</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.259</td>
<td>0.003</td>
<td>0.899</td>
<td>0.100</td>
<td>0.001</td>
<td>0.986</td>
<td>0.012</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AIC</td>
<td>0.200</td>
<td>0.002</td>
<td>0.963</td>
<td>0.037</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SIC</td>
<td>0.270</td>
<td>0.008</td>
<td>0.963</td>
<td>0.037</td>
<td>0.000</td>
</tr>
<tr>
<td>T = 300</td>
<td>2</td>
<td>0.783</td>
<td>0.404</td>
<td>0.905</td>
<td>0.096</td>
<td>0.000</td>
<td>0.914</td>
<td>0.086</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.384</td>
<td>0.025</td>
<td>0.882</td>
<td>0.118</td>
<td>0.000</td>
<td>0.974</td>
<td>0.026</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.238</td>
<td>0.001</td>
<td>0.868</td>
<td>0.132</td>
<td>0.000</td>
<td>0.981</td>
<td>0.019</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AIC</td>
<td>0.174</td>
<td>0.000</td>
<td>0.959</td>
<td>0.041</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SIC</td>
<td>0.224</td>
<td>0.002</td>
<td>0.959</td>
<td>0.041</td>
<td>0.000</td>
</tr>
</tbody>
</table>

See footnotes to Table 6.
<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>VAR(1)-(MA1)</th>
<th>VAR(1)-(MA2)</th>
<th>VAR(3)-(MA1)</th>
<th>VAR(3)-(MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.064</td>
<td>0.064</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 2</td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 3</td>
<td>2</td>
<td>0.983</td>
<td>0.757</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>0.218</td>
<td>0.081</td>
<td>0.192</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>2</td>
<td>0.195</td>
<td>0.020</td>
<td>0.195</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Lag/sample</th>
<th>CI rank</th>
<th>AIC</th>
<th>SIC</th>
<th>AIC</th>
<th>SIC</th>
<th>AIC</th>
<th>SIC</th>
<th>AIC</th>
<th>SIC</th>
<th>AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T = 200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.085</td>
<td>0.073</td>
<td>0.147</td>
<td>0.013</td>
<td>0.202</td>
<td>0.058</td>
<td>0.142</td>
<td>0.010</td>
<td>0.142</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.992</td>
<td>0.998</td>
<td>0.810</td>
<td>0.995</td>
<td>0.723</td>
<td>0.994</td>
<td>0.590</td>
<td>0.994</td>
<td>0.590</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.068</td>
<td>0.068</td>
<td>0.160</td>
<td>0.007</td>
<td>0.205</td>
<td>0.051</td>
<td>0.149</td>
<td>0.005</td>
<td>0.149</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.947</td>
<td>1.00</td>
<td>0.981</td>
<td>1.00</td>
<td>0.981</td>
</tr>
<tr>
<td>Lag = 1</td>
<td>1</td>
<td>0.069</td>
<td>0.069</td>
<td>0.002</td>
<td>0.002</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T = 300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 2</td>
<td>1</td>
<td>0.194</td>
<td>0.079</td>
<td>0.171</td>
<td>0.021</td>
<td>0.208</td>
<td>0.035</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.998</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>T = 300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 3</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T = 300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = 4</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T = 300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag = AIC</td>
<td>1</td>
<td>0.083</td>
<td>0.069</td>
<td>0.142</td>
<td>0.022</td>
<td>0.180</td>
<td>0.059</td>
<td>0.118</td>
<td>0.011</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.999</td>
<td>1.00</td>
<td>0.902</td>
<td>1.00</td>
<td>0.918</td>
<td>0.999</td>
<td>0.734</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lag = SIC</td>
<td>1</td>
<td>0.069</td>
<td>0.069</td>
<td>0.140</td>
<td>0.007</td>
<td>0.201</td>
<td>0.058</td>
<td>0.116</td>
<td>0.00</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.966</td>
<td>1.00</td>
<td>0.997</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*See footnotes to Table 5.*

*All estimations are based on the assumption that it is known that the data are generated according to a VEC.*
<table>
<thead>
<tr>
<th>Sample</th>
<th>Lag</th>
<th>VAR(1)-(MA1) AIC</th>
<th>SIC</th>
<th>VAR(1)-(MA2) AIC</th>
<th>SIC</th>
<th>VAR(3)-(MA1) AIC</th>
<th>SIC</th>
<th>VAR(3)-(MA2) AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.341</td>
<td>0.115</td>
<td>0.554</td>
<td>0.349</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.270</td>
<td>0.040</td>
<td>0.358</td>
<td>0.078</td>
<td>0.296</td>
<td>0.043</td>
<td>0.602</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>0.344</td>
<td>0.113</td>
<td>0.282</td>
<td>0.018</td>
<td>0.253</td>
<td>0.022</td>
<td>0.240</td>
<td>0.005</td>
</tr>
<tr>
<td>SIC</td>
<td></td>
<td>0.341</td>
<td>0.115</td>
<td>0.353</td>
<td>0.072</td>
<td>0.304</td>
<td>0.051</td>
<td>0.320</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.301</td>
<td>0.080</td>
<td>0.563</td>
<td>0.329</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.241</td>
<td>0.022</td>
<td>0.360</td>
<td>0.053</td>
<td>0.283</td>
<td>0.028</td>
<td>0.617</td>
<td>0.288</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>0.308</td>
<td>0.078</td>
<td>0.193</td>
<td>0.002</td>
<td>0.208</td>
<td>0.007</td>
<td>0.160</td>
<td>0.00</td>
</tr>
<tr>
<td>SIC</td>
<td></td>
<td>0.301</td>
<td>0.080</td>
<td>0.270</td>
<td>0.019</td>
<td>0.271</td>
<td>0.026</td>
<td>0.218</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.310</td>
<td>0.057</td>
<td>0.599</td>
<td>0.342</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.224</td>
<td>0.011</td>
<td>0.352</td>
<td>0.043</td>
<td>0.296</td>
<td>0.015</td>
<td>0.687</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>0.307</td>
<td>0.054</td>
<td>0.165</td>
<td>0.001</td>
<td>0.191</td>
<td>0.002</td>
<td>0.139</td>
<td>0.00</td>
</tr>
<tr>
<td>SIC</td>
<td></td>
<td>0.310</td>
<td>0.057</td>
<td>0.220</td>
<td>0.006</td>
<td>0.230</td>
<td>0.005</td>
<td>0.179</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*See footnotes to Table 5

*All estimations are based on the assumption that it is known that the data are generated according to a levels VAR.
<table>
<thead>
<tr>
<th>Sample</th>
<th>Lag</th>
<th>VAR(1)-(MA1)</th>
<th>VAR(1)-(MA2)</th>
<th>VAR(3)-(MA1)</th>
<th>VAR(3)-(MA2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AIC</td>
<td>SIC</td>
<td>AIC</td>
<td>SIC</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.829</td>
<td>0.567</td>
<td>0.995</td>
<td>0.937</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>———-</td>
<td>———-</td>
<td>———-</td>
<td>———-</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.828</td>
<td>0.609</td>
<td>0.968</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.00</td>
<td>0.00</td>
<td>0.994</td>
<td>0.924</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.978</td>
<td>0.828</td>
<td>1.00</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>———-</td>
<td>———-</td>
<td>———-</td>
<td>———-</td>
</tr>
<tr>
<td>1</td>
<td>AIC</td>
<td>0.930</td>
<td>0.512</td>
<td>0.997</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>SIC</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.996</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>———-</td>
<td>———-</td>
<td>———-</td>
<td>———-</td>
</tr>
<tr>
<td>AIC</td>
<td>0.963</td>
<td>0.870</td>
<td>1.00</td>
<td>0.930</td>
<td>0.787</td>
</tr>
<tr>
<td>SIC</td>
<td>———</td>
<td>———</td>
<td>1.00</td>
<td>1.00</td>
<td>0.761</td>
</tr>
</tbody>
</table>

\(^a\)See Footnotes to Table 6b
\(^b\)All estimations are based on the assumption that it is known that the data are generated according to a VAR in differences.
Causality and Predictive Accuracy Tests 143

results did not depend on the specification of \( l \), however, and so in Table 1 we report results only for the case where \( l = 0 \).

Sequential Wald test results are given in Tables 2–4. Single lag cases are not reported for \( \lambda_{p3} \), as the test is not applicable in this case. In these tables, \( lag = 1 \) corresponds to zero lags when the data are differenced, so that the corresponding error-correction models used to calculate \( \lambda_{p1} \), and \( \lambda_{p2} \) have no short-run dynamics. In general, this has no effect on our results, as the DGPs are designed so as to impact causality both through the short-run dynamics (if there is any) as well as through the long-run dynamics (the error-correction part). In most cases, the best size tests are associated with one lag and with cointegrating rank either (1) selected to be the truth (one) or (2) selected using a Johansen likelihood ratio trace test without a constant. In fact, the very similar sizes for the two “good” cointegrating (CI) rank cases arise because the Johansen method is usually able to “select” the true CI rank, even in the presence of substantial model complexity [i.e., for VAR(3) models with an MA component]. However, of note is that when the (CI-1)–VAR(3)–MA2 results are examined (see Table 2), the \( \lambda_{p1}–\lambda_{p3} \) sizes range from “bests” of 0.194–0.234 (\( T = 100 \)); 0.118–0.161 (\( T = 200 \)); and 0.099–0.149 (\( T = 300 \)), when lags are selected using the AIC, and are always worse when lags are selected using the SIC. Interestingly, for the CI-2 case (see Table 3), SIC lag selection leads to empirical sizes closer to nominal in the VAR(3)–MA(2) case, suggesting that the evidence in favor of using the AIC for more complex models is not always clearcut. Indeed, for corresponding \( \lambda_{LW} \) tests, size based on the SIC lag selection criterion is usually better than that based on the AIC criterion, and is: 0.130 (\( T = 100 \)), 0.109 (\( T = 200 \)), and 0.090 (\( T = 300 \)) for CI-1. In fact, the only DGPs for which AIC is preferred over SIC for lag selection are VAR(3)–MA2 specifications. This again underscored the usefulness of the AIC only in more complex models.

The \textit{ex ante} forecasting based model selection type approaches perform quite well (see Tables 9–11, 13, and 14). In particular, for complex DGPs [e.g., VAR(3)–MA2] the reported sizes of the MSE, MAD, and MAPE tests are very small (from 0.002 to 0.015) for lags selected using either the AIC or the SIC. As is generally the case, though, lag selection based on the SIC is dominated by selection based on the AIC for more complex models. Even given the clear success of the \textit{ex ante} approaches, it should be noted that our experimental design is very simple. In addition, there is much research currently underway in the field of econometrics into the construction and performance of out-of-sample predictive ability tests, so that much remains to be learned.

The in-sample model selection approaches fare very well also (see Tables 5–8 and 12–15). In particular, the SIC seems to provide a very
reasonable alternative to standard hypothesis testing, when the null is Granger noncausality. In Tables 5 and 10 note that the number of “wins” (based on a comparison of model selection criteria for the different models) for a given “true” DGP and for each of the estimated cointegrating, difference stationary, and levels stationary models are reported. In this way, we can get an idea as to which type of model is preferred by AIC and SIC. As mentioned above, the way these numbers are reported is as follows. From amongst all lag and CI rank combinations, the AIC and SIC-best models are used in each simulation to “test” for noncausality. If the AIC-best model is a cointegration model, then a “win” is tallied for “cointegration.” if the AIC-best model is a difference model, then a “win” is tallied for “differences,” and so on. The rest of the tables on in-sample model selection report rejection frequencies. Notice in these tables that the AIC and SIC tend to select the “true” DGP (e.g., VEC models are found to be preferred whenever data are generated according to VECs). Even casual inspection of the results suggests that the SIC approach for checking Granger causality performs better than the AIC. In addition, both model selection criteria are somewhat sensitive to the combination of DGP and estimation method. For example, results are very poor when data are generated according to a VEC, but levels or difference stationary models are estimated, and fixing the CI rank at 2 when it is actually 1 also results in poor performance. Thus, although this sort of model selection appears promising, there do appear to be frailties with this approach (as with all approaches when used with mis-specified models).

4.2 Difference Stationary Models

In order to examine the performance of Wald tests directly using differenced data ($\lambda_{DS}$), we also generated a number of difference stationary DGPs, and applied all seven types of causality tests. The results are omitted here for the sake of brevity, but are reported in Swanson et al., and Pisu (SOP: 1996, Tables 17–23). Of note is that the test based on $\lambda_{DS}$ is very precise for VAR(1) ($T = 100, 200, 300$), and for the VAR (3) ($T = 300$), when the correct number of lags is used (e.g., the sizes are 0.054, 0.05, 0.052, and 0.055, respectively). However, when an MA component (MA2) is introduced, the standard difference stationary type Wald tests are dominated by the levels type Toda and Yamamoto tests for all lags, as well as all sample sizes for the VAR(1) case. For example, for VAR(1)–MA2 sizes for the $\lambda_{DS}$ test based on lags selected using the AIC are 0.176, 0.111, and 0.095, for $T = 100, 200$, and 300, respectively. These are the “best” sizes available for this case. The corresponding values associated with the $\lambda_{LH}$ type test are 0.122, 0.099, and 0.080. In addition, it is worth
noting that sequential Wald tests yield rather uninspired results when faced with difference stationary data (see Tables 18–21 in SOP). Indeed, in most cases [except for VAR(3)–MA DGPs] the sizes of $\lambda_{p1}$–$\lambda_{p3}$ are in the 1–20% range.

Findings based on the use of in-sample model selection procedures are qualitatively the same as when VEC DGPs are generated. However, the SIC-type causality test seems to perform seven better, when lags are chosen using the AIC. This may be because AIC infrequently picks too few lags for inclusion in the estimated VAR model.

5 CONCLUSIONS

In this chapter we have discussed and compared five different Wald-type tests of the null hypothesis of noncausality, as well as two different tests based on model selection approach to assessing the predictive ability of the variable(s) whose causal effect is of interest. Our main purpose was to gather together and discuss a reasonable number of recent tests, some of which are asymptotically valid under the null hypothesis regardless as to whether or not the data are integrated of order zero, integrated of higher order, or cointegrated of some order. We also examined a number of model selection based approaches to assessing causal ordering, including those based on in-sample complexity penalized likelihood criteria, and those based on the examination of ex ante forecasting ability. Finally, we obtained results that have some implications for empirical work. For example, our examination suggests that three types of tests perform reasonably well (in terms of size), regardless as to the integratedness and cointegratedness properties of the variables in the system, and regardless as to whether or not the model is correctly specified. The first is the standard Wald-type test, which is based on levels regressions, and which adds “surplus” lags to account for integratedness and cointegratedness. This test is due to Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996). It should be noted, however, that the local power properties of such tests may indeed be suspect, although concrete results on this matter have yet to be published. The second consists of the model selection based approach of using the Schwarz information criterion to select a “best” model, and then either “reject” or “accept” the noncausality null based on whether the causal variable of interest appears in the “best” model. In is perhaps worth stressing that model selection type approaches focus more on predictive ability, and as such may be particularly useful when the issue at hand is forecastability. In-sample hypothesis testing, on the other hand, may be preferable when the statistical significance of particular coefficients in a model is the issue being examined (e.g., when testing economic theories).
ACKNOWLEDGMENTS

The authors wish to thank Lutz Killian, seminar participants at the 1997 Winter Meetings of the Econometric Society, and seminar participants at the University of Pennsylvania for useful comments on an earlier version of the paper on which this chapter is based. Swanson thanks the Private Enterprise Research Center at Texas A&M University for financial support.

REFERENCES


Finite Sample Performance of the Empirical Likelihood Estimator Under Endogeneity

Ron C. Mittelhammer
Washington State University, Pullman, Washington, U.S.A.

George G. Judge
University of California at Berkeley, Berkeley, California, U.S.A.

1 INTRODUCTION

There is a growing body of evidence that traditional asymptotically efficient moment-based estimators for the linear structural model may have large biases for the relatively small sample sizes usually encountered in applied economic research (see, e.g., Newey and Smith, 2000). Furthermore, instrumental variables that are utilized in the context of a set of moment–orthogonality conditions may be only weakly correlated with the endogenous variables in the model and consequently the parameters of the structural model are only poorly or weakly identified. In these situations it is generally recognized that, for both traditional and nontraditional estimators, bias and variability problems may arise and that large sample normal approximations provide a poor basis for finite sample performance (see, e.g., Nelson and Startz, 1990; Maddala and Jeong, 1992; Bound et al., 1995; Stock and Wright, 2000). Given these general problem areas the focus of this chapter is on semiparametric estimation and inference relative to the response parameters for a linear structural statistical model. Using Monte
Carlo sampling procedures and a range of underlying data-sampling processes, we provide finite sample comparisons of the two-stage least squares (2SLS) and the empirical likelihood (EL) estimator for recovering the unknown parameters of the linear structural statistical model.

In finite sample situations where the instrumental variables are either well or weakly correlated with the endogenous variable(s) in question and given that the orthogonal moment conditions hold, we seek answers to the following questions:

1. Do empirical likelihood (EL) estimators exhibit substantial small sample bias and, in terms of bias, is the EL estimator superior to traditional estimators?
2. In terms of precision is the EL estimator superior to traditional estimators?
3. In terms of inference relating to structural model parameters, do traditional testing procedures, relative to EL testing procedures, lead to more accurate coverage but wider confidence intervals?
4. What is the size performance of the traditional and EL inference procedures relative to testing the validity of the moment restrictions?

The format of the chapter is as follows: In Section 2 the linear structural statistical model is defined and the competing estimators are specified. In Section 3 the design of the sampling experiment is presented and the alternative data-sampling processes are defined. In Section 4 the sampling results as they relate to estimation and inference are reported. The implications of the results are summarized in Section 5. Some supplementary information in table form is presented in the Appendix.

2 STATISTICAL MODELS AND ESTIMATORS

Consider a single equation in a system of equations that has the semiparametric linear statistical response model form $y = X\beta + \varepsilon$, where we observe a vector of sample observations $y = (y_1, y_2, \ldots, y_n)$, $X$ is a $(n \times k)$ matrix of realizations of stochastic variables, $\varepsilon$ is an unobservable random noise vector with mean 0 and covariance matrix $\sigma^2 I_n$, and $\beta \in B$ is a $(k \times l)$ vector of unknown parameters. If one or more of the regressors is correlated with the equation noise, then $E[n^{-1}X'\varepsilon] \neq 0$ or $\text{plim } [n^{-1}X'\varepsilon] \neq 0$ and traditional Gauss–Markov procedures such as the least squares (LS) estimator, or equivalently the method of moments (MOM)–extremum estimator $\hat{\beta}_{\text{mom}} = \arg \text{pe} \{n^{-1}X'(y - X\beta) = 0\}$, are biased and inconsistent, with unconditional expectation and probability limit given by $E[\hat{\beta}] \neq \beta$ and $\text{plim } [\hat{\beta}] \neq \beta$. 
Given a sampling process characterized by nonorthogonality of $X$ and $\varepsilon$, to avoid the use of strong underlying distributional assumptions it is conventional to introduce additional information in the form of independent realizations of a $(n \times m), m \geq k$ random matrix $Z$ of instrumental variables, whose elements are correlated with $X$ but uncorrelated with $\varepsilon$. This information is introduced into the statistical model by using a sample analog to the orthogonality of instruments and model noise as

$$ h(y, X, Z; \beta) = n^{-1}[Z(y - X\beta)] \xrightarrow{p} o $$

If $m = k$ the sample moments are solved for the basic instrumental variable (IV) estimator $\hat{b}_n = (Z'X)^{-1}Z'Y$, which is consistent and asymptotically normally distributed (Mittelhammer et al., 2000). When $m > k$ the 2SLS estimator, $\beta_{2sls} = (X'P_zX)^{-1}X'P_zY$, where $P_z = Z(Z'Z)^{-1}Z'$ is the projection matrix based on $Z$, and the GMM estimator, are popular choices with optimal asymptotic properties.

### 2.1 The MEL Estimator

Alternatively, the maximum empirical likelihood (MEL) approach (Owen, 1988, 1991; Qin and Lawless, 1994; Mittelhammer et al., 2000) offers another way to seek a unique solution for $\beta$ using the instrumental variable-based moment information in Section 2.1. Under the MEL concept, empirical likelihood weights based on multinomial distributions supported on a sample of observed values are used to reduce the infinite dimensional problem of nonparametric likelihood estimation to a finite dimensional one. In an instrumental variable context for the linear structural equation, this leads to the following extremum-type estimator for the parameter vector $\beta$ that defines the MEL estimator $\hat{\beta}_{EL}$ as

$$ \hat{\beta}_{EL} = \arg \max_{\beta} \ell_{E}(\beta) $$

where

$$ \ell_{E}(\beta) \equiv \max_{w} \left\{ n^{-1} \sum_{i=1}^{n} \ln(w_i) \left| \sum_{i=1}^{n} w_i x_i \right| (y_i - x_i \beta) = 0, \right. $$

$$ \left. \sum_{i=1}^{n} w_i = 1, w_i \geq 0 \ \forall i, \ \beta \in B \right\} $$

(3)

can be interpreted as a semiparametric profile empirical likelihood for $\beta$ (Murphy and Van Der Vaart, 2000). If this Owen MEL criterion is used, the estimation objective involves finding the feasible weights $\hat{w}$ that maximize the empirical probabilities assigned to the observed set of observations.
The Lagrange function associated with the empirical likelihood formulation is

\[ L(w, \alpha, \eta) = n^{-1} \sum_{i=1}^{n} \ln(w_i) - \alpha' \sum_{i=1}^{n} w_i z'_i (y_i - x_i \beta) - \eta \left( \sum_{i=1}^{n} w_i - 1 \right) \]  

(4)

and the optimal \( w_i \) can be expressed in terms of \( \beta \) and \( \alpha \) as

\[ w_i(\alpha, \beta) = \left[ n(\alpha' z'_i (y_i - x_i \beta) + 1) \right]^{-1} \]  

(5)

where \( \alpha \) is the \( m \times 1 \) optimal Lagrange multiplier vector and \( \beta \) solves Eq. (2). The EL estimator \( \hat{\beta}_{\text{EL}} \) is consistent, asymptotically normally distributed, and asymptotically efficient. A consistent empirical representation of the asymptotic distribution is given by

\[ \hat{\beta}_{\text{EL}} \overset{d}{\sim} N \left( \beta, n^{-1} \left( \sum_{i=1}^{n} \hat{w}_i X_i Z_i \right) \left[ \sum_{i=1}^{n} \hat{w}_i \hat{e}^2_i Z_i Z_i \right]^{-1} \left( \sum_{i=1}^{n} \hat{w}_i X'_i Z_i \right) \right)^{-1} \]  

(6)

or equivalently,

\[ \hat{\beta}_{\text{EL}} \overset{d}{\sim} N \left( \beta, n^{-1} \left( X'(Z \odot \hat{w}) \right) \right. \]

\[ \left. \times \left[ (Z \odot \hat{w}^{1/2} \odot \hat{\epsilon}) (Z \odot \hat{w}^{1/2} \odot \hat{\epsilon}) \right]^{-1} (Z \odot \hat{w}) X \right)^{-1} \]  

(7)

where \( \odot \) denotes the extended Hadamard (elementwise) product, and \( \hat{\epsilon} = Y - X \hat{\beta}_{\text{EL}} \). The \( \hat{w} \) vector in Eq. (6) or (7) can be replaced by the vector \( n^{-1}1 \) to define an alternative asymptotically valid representation of the covariance matrix of the EL estimator.

Calculating the solution to the MEL estimation problem will generally require that a computer-driven maximization algorithm be employed. When \( m = k \), the solution to the EL extremum problems leads back to the standard IV estimator \( \hat{\beta}_n \) with \( \hat{w}_i = n^{-1} \). When \( m > k \), the estimating equations over determine the unknown parameter values to be recovered and a nontrivial EL solution results. Computing solutions to EL-constrained optimization problems can present formidable numerical challenges. This computing problem arises because, in the neighborhood of the solution to such problems, the gradient matrix associated with the moment constraints approaches an ill-conditioned state of being less than full rank. This occurs by design in these types of problems because the fundamental method by which the EL method addresses the overdetermined nature of the moment...
conditions is by finding EL weights $\hat{w}_i$, for $i = 1, \ldots, n$, that ultimately transform the moment equations into a functionally dependent, lower rank (equal to $k < m$) system of equations capable of being solved uniquely for the parameters. This creates instability in gradient-based constrained optimization algorithms regarding the representation of the feasible spaces and feasible directions for such problems.

2.2 MEL Inference

In terms of inference the EL function $\ell_E(\beta)$ bears a strong analogy to the likelihood function used in traditional ML procedures for developing tests and confidence regions that are defined in terms of likelihood ratios. In this context, the empirical likelihood ratio (ELR) for the test of the hypothesis $\beta = \beta_0$, or more generally for testing the linear combinations hypothesis $c\beta = r$, is given by [recall Eq. (2)]

$$LR_{EL}(y) = \frac{\max_{\beta}[\ell_E(\beta) \text{s.t. } c\beta = r]}{\max_{\beta} \ell_E(\beta)}$$  \hspace{1cm} (8)$$

and

$$-2\ln(LR_{EL}(Y))$$ \hspace{1cm} (9)

under $H_0$ when $m \geq k$. One can also base tests of $c\beta = r$ on the Wald criterion in the usual way utilizing the inverse of the asymptotic covariance matrix exhibited in Eq. (6) or (7) as the weight matrix of a quadratic form in the vector $c\hat{\beta}_{EL} = -r$, as well as construct tests based on the Lagrange multipliers associated with the constraints $c\beta = r$ imposed on the EL optimization problem, Eq. (2) or (3).

Interest may also center on testing the validity of the moment equations, the null hypothesis then being $H_0: E[h(\mathbf{y}, \mathbf{X}, \mathbf{Z}\beta)] = \mathbf{0}$. A natural way to make this test is to use an ELR statistic based on a ratio of empirical likelihoods for which the EL weights are either constrained to satisfy the moment conditions (the numerator of the ratio) or else are unconstrained by the moment conditions, in which case the empirical likelihood is based on $w_i = n^{-1}, \forall i$. If the moment constraints are correct, the ratio is distributed asymptotically as a central chi square, with degrees of freedom equal to $m - k$, $m$ being the number of functionally independent moment conditions and $k$ being the number of unknown parameters involved in the specification of the moment conditions. Details of the empirical implementation of tests used in the Monte Carlo experiments contained in this chapter are presented in Section 4. For additional discussion of the ELR and other tests, see Chapters 12 and 13 in Mittelhammer et al. (2000).
2.3 Remark on Performance Evaluations

Although the estimators are asymptotically equivalent, and are also equivalent in finite samples if \( m = k \), when \( m > k \) the sampling properties of the 2SLS and EL estimators can differ in finite samples. Since each of the estimators involves a bias and a variance component, this raises the question of how to evaluate estimator performance. In some cases a researcher may be willing to accept some degree of bias if it improves precision in estimation and prediction. In situations such as this where bias–variance trade offs are being considered, the squared error loss (SEL) measure \( L(\beta, \hat{\beta}) = \| \beta - \hat{\beta} \|^2 \) or \( L(y, X\hat{\beta}) = \| y - x\hat{\beta} \|^2 \) provides a basis for comparing the performance of the two competing estimators and this is one basis we use for comparing the performance of the estimators in Section 4.

3 DESIGN OF SAMPLING EXPERIMENTS

In terms of the EL-type formulation of Section 2, the solution for the optimal weights cannot be expressed in closed form. Moreover, the finite sample probability distributions of the traditional 2SLS estimator are generally intractable. Consequently, the finite sample properties of the aforementioned estimation procedures cannot be derived from direct evaluation of the estimator’s functional form. Therefore, we use a Monte Carlo sampling experiment to identify and compare the repeated sampling performance of the estimators in terms of the expected squared error loss associated with estimating \( \beta \) and predicting \( y \). We also provide results relating to the average bias and variances associated with the competing estimators. While these results are specific to the collection of particular Monte Carlo experiments analyzed, the extensive sampling evidence reported does provide an indication of the types of relative performance that can occur over a range of well and weakly identified data-sampling scenarios.

3.1 Experimental Design

Consider a sampling process of the following form:

\[
Y_{i1} = Z_{i1}\beta_1 + Y_{i2}\beta_2 + \varepsilon_i = X\beta + \varepsilon_i
\]  
\[
Y_{i2} = \sum_{j=1}^{4} \pi_j Z_{ij} + v_i = Z_{i}\pi + v_i
\]

where \( X_i = (Z_{i1}, Y_{i2}) \) and \( i = 1, 2, \ldots, n \). The two-dimensional vector of unknown parameters, \( \beta \), in Eq. (10) is arbitrarily set equal to the vector \([-1, 2]^T\). The outcomes of the \((6 \times 1)\) random vector \([Y_{i2}, \varepsilon_i, Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]\) are generated \( iid \) from a multivariate normal distribution with a zero mean.
vector and unit variances for all random variables other than $\varepsilon$, which has a variance of 4. Also, there are various other conditions relating to the correlations existent among the six scalar random variables. The values of the $p$'s in Eq. (11) are clearly determined by the regression function between $Y_2$ and $[Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]$, which is itself a function of the covariance specification relating to the marginal normal distribution associated with the $(5 \times 1)$ random vector $[Y_2, Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]$. Thus, the $p$'s generally change as the scenario postulated for the correlation matrix of the sampling process changes. In this sampling design, the outcomes of $[Y_2, Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]$ are then calculated by applying Eqs. (10) and (11) to the outcomes of $[Y_2, Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]$.

3.2 Monte Carlo Scenario Characteristics

Regarding the details of the sampling scenarios simulated for this set of Monte Carlo (MC) experiments, the focus was on sample sizes of $n = 25, 50, 100,$ and $250$. The outcomes of $\varepsilon_i$ were generated independently of the vector $[Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]$ so that the correlations between $\varepsilon_i$ and the $Z_{ij}$'s were zeros, thus fulfilling a fundamental condition for $[Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}]$ to be considered a set of valid instrumental variables for estimating the unknown parameters in Eq. (10). Regarding the degree of nonorthogonality–identifiability in Eq. (10), a correlation of 0.75 between the random variables $Y_2$ and $\varepsilon_i$ was utilized to simulate a relatively strong correlation–nonorthogonality relationship between the explanatory variable $Y_2$ and the equation noise $\varepsilon_i$. For each sample size, alternative scenarios were examined relating to both the degree of correlation existing between each of the instruments in the matrix $Z$ and the $Y_2$ variable, and the levels of collinearity existing among the instrumental variables themselves. By varying the degrees of intercorrelation among the variables, the overall correlation of the instrumental variables with $Y_2$ is effected, and contributes to determining the overall effectiveness of the set of instruments in predicting values of the endogenous $Y_2$.

The major characteristics of each of the scenarios are delineated in Table 1. In general, the scenarios range from very weak but independent instruments to stronger but highly collinear instruments. All told, there were 10 different MC experimental designs in combination with the four different samples, resulting in 40 different sampling scenarios in which to observe estimator and inference behavior. For a useful summary of experimental designs used by others to represent weakly determined situations, see Zivot (2001).

The reported sampling properties relating to estimation objectives are based on 2500 MC repetitions, and include estimates of the empirical risks,
based on a SEL measure, associated with estimating $\beta$ with $\hat{\beta}$ (parameter estimation risk) and estimating $y$ with $\hat{y}$ (predictive risk). We also report on the average estimated bias in the estimates, $\text{bias}(\hat{\beta}) = E[\hat{\beta}] - \beta$, and the average estimated variances of the estimates, $\text{var}(\hat{\beta})$.

Overall, three different estimators were examined including the 2SLS estimator, the least squares (LS) estimator, and the EL estimator. Estimator outcomes were obtained using GAUSS software, and in particular, the EL-type solutions were calculated using the GAUSS-constrained optimization application module provided by Aptech Systems, Maple Valley, WA, U.S.A.

4 SAMPLING RESULTS

The sampling designs specified in Section 3 range from well-behaved data-sampling processes, involving instrumental variables that have the recommended correlation structure for which the traditional 2SLS estimator would be expected to perform reasonably well, to other extremes involving low correlation of the IVs with the included endogenous variable and weak identification, high multicollinearity, and poorer fitting models. In reporting the sampling results we discuss the

\[ \beta = [-1.2]' \] and $\sigma^2 = 4$ for scenarios 1–10, and $\sigma^2_{\epsilon_i} = \sigma^2_{\epsilon_j} = 1$, $\forall i$ and $j$. $\rho_{\epsilon, \epsilon_i}$ denotes the correlation between $Y_2$ and $e$, and measures the degree of nonorthogonality; $\rho_{\epsilon, \epsilon_j}$ denotes the correlation between $Y_2$ and $Z_i$; $\rho_{\epsilon, \epsilon_j}$ denotes the common correlation between $Y_2$ and each of the three instrumental variables not included in the first equation; $\rho_{\epsilon, \epsilon_j}$ denotes the common correlation between $Z_1$ and the remaining three $Z_i$‘s; $\rho_{\epsilon, \epsilon_j}$ denotes the common correlation between the three instrumental variables not included in the first equation; $\rho_{\epsilon, \epsilon_j}$ denotes the population correlation between $Y_2$ and $Y_2 = Z_1$; and $R^2_{Y_1, Y_2}$ denotes the population squared correlation between $Y_1$ and $Y_1 = X\beta$.
empirical evidence regarding the performance of the 2SLS and EL estimators for the well- and ill-behaved sampling models for samples sizes of $n = 25, 50, 100$, and 250. In each case we report estimation sampling performance relating to empirical mean parameter squared error loss (SEL), mean squared prediction error (MSPE), bias, and variance. In terms of inference, we report sampling performance relating to test sizes for tests of moment validity, as well as coverage and power observations for tests on structural model parameters. To conserve space and facilitate discussion, some of the sampling results are reported in graphical form, and we generally defer observations on LS estimator performance to the results tables. Appendix tables contain detailed information underlying the graphs and contain sampling results for alternative sample sizes and the bias and variance sampling results for $\beta_1$ and $\beta_2$.

### 4.1 Discussion

#### 4.1.1 Empirical Parameter SEL Performance

The well-behaved data-sampling processes that appear strongly identified and that involve low correlations among the instruments, moderate to high correlations between $Y_2$ and the $Z_j$’s, and low or moderate correlation between $Y_2$ and $Z_1$ are represented by data-sampling processes (DSPs) defined by scenarios 1, 2, 3, and 4. In Figure 4.1 the empirical parameter

---

**FIGURE 1** Parameter SEL for selected DSPs.
estimation risks for these DSPs and the 2SLS and EL estimators are presented for samples sizes of \( n = 50 \) and 100. In Fig. 1, for sample sizes \( n = 50 \) and 100, the 2SLS estimator is seen to be superior in empirical mean parameter SEL for all four scenarios, but the empirical risk differences are small. For \( n = 250 \), the 2SLS estimator continues to exhibit empirical risk dominance, but at this sample size the nominal levels of risk are small for either estimator, with the empirical risk differences between the two estimators also being relatively small. For \( n = 25 \) the 2SLS remains risk superior in the first three scenarios, but EL is superior in the fourth, the risk differences between the estimation procedures again being relatively small.

The ill-behaved, weakly identified DSPs are defined by scenarios 5–10. These sampling processes depart from the a priori IV ideal type of correlation structure and are consistent with sampling processes found in practice where the instruments have low correlation with the individual endogenous variable, and the parameters are more poorly identified. The empirical parameter risks of the 2SLS and EL estimators for these sampling designs are depicted in Fig. 1 for sample sizes \( n = 50 \) and 100. In the set of graphed comparisons the EL estimator is superior in three-quarters of the scenarios, where only in scenario 10 were there any effective risk gains from using 2SLS in place of EL. The results in Table A.1 indicate that the same relative risk behavior occurs when \( n = 5 \), while as \( n \) increases to 250 the performance of the 2SLS estimator improves relative to the EL estimator, as does the overall performance of both estimators. Overall, the differences between the two estimators in empirical mean parameter SEL is generally small, with the 2SLS estimator being a marginally better choice in well-behaved sampling processes characterized by moderate-to-strong parameter identification, and the EL estimator being marginally superior in cases of relatively poorly behaved sampling processes in which there is higher collinearity, weaker correlation between instruments and \( Y_2 \), and relatively weak parameter identification.

### 4.1.2 Empirical Bias

The empirical bias in estimating \( \beta_2 \) for \( n = 50 \) and 100 is presented in Fig. 2, and the biases for both \( \beta_1 \) and \( \beta_2 \), as well as results for sample sizes \( n = 25 \) and 250, are given in Table A.2. In the case of the well-behaved sampling designs 1–4, and for \( n \geq 50 \), the 2SLS estimator dominates the EL estimator in terms of having a smaller \( \beta_2 \) bias, and the relative degree of dominance increases as \( n \) increases. While the empirical bias of each estimator decreases as \( n \) increases, the bias of the EL estimator is more
persistent and was still 5 to 8% of the true $\beta_2$ value for $n = 250$. As expected, the LS estimator exhibited a large relative bias and this bias did not decrease as $n$ increased. Interestingly, for the smallest sample size of $n = 25$, the EL estimator dominates the 2SLS estimator in terms of $\beta_2$ bias across all four scenarios.

In terms of the empirical bias in estimating $\beta_2$ for the ill-behaved sampling scenarios 5–10, both estimators have substantial bias. For $n = 25$ the EL estimator has slightly better $\beta_2$-bias performance than the 2SLS estimator. As sample sizes increase, the biases of both estimators decrease, but the 2SLS estimator tends to be slightly better than the EL estimator in the majority of cases. For the severely ill-behaved scenarios 7 and 8 in which the instruments are extremely poor in predicting outcomes of $Y_2$, there is no perceptible decrease in estimator bias as the sample size increases. Overall, except in the case of both very small sample sizes and ill-behaved data-sampling processes, the 2SLS estimator is generally marginally better than the EL estimator in terms of $\beta_2$ bias.

The comparison between 2SLS and EL is not as clear cut when comparing $\beta_1$-bias results, where EL is superior in more cases, and EL superiority is not isolated to the smallest sample size but rather occurs in at least half of the cases for each of the four sample sizes. However, the large majority of the bias levels for the $\beta_1$ parameter are nominally quite small, and the bias differences are of generally little practical consequence.
4.1.3 Empirical Variance

For scenarios 1–4, the empirical variance of the 2SLS estimator for $\beta_2$ is smaller than the EL estimator in a majority of the cases. However, the size of the variance differences is small. As $n$ increases the sampling precision of both of the estimators increases (see Fig. 3 and Table A.3).

In terms of the empirical $\beta_2$ variance for scenarios 5–10, as sample size increases the empirical variances decrease (see Fig. 3). However, similar to the case of bias, the rate of decline in variance is markedly slower for the severely ill-behaved scenarios 7 and 8. As sample size increases, the EL estimator tends to dominate the 2SLS estimator empirically in terms of achieving a smaller estimator $\beta_2$ variance.

In the case of $\beta_1$-variance comparisons, the 2SLS is superior in the large majority of cases. However, the $\beta_1$ variance of each estimator is generally much smaller than for $\beta_2$, and the variance differences between the estimators are generally small.

Considering variance and bias characteristics simultaneously, the marginally superior parameter SEL performance of EL for the ill-behaved DSPs can generally be attributed to reductions in the variances of the estimators of the structural parameters, and generally not to superior bias characteristics of the EL estimator.

FIGURE 3 Empirical estimator variances for selected DSPs.
4.1.4 Empirical Prediction Errors

In terms of mean squared prediction error, the EL estimator is superior to the 2SLS estimator in the large majority of cases, and the frequency and level of superiority increases as \( n \) increases (see Table A.1). The EL estimator dominates the 2SLS estimator in average squared prediction risk when \( n \geq 100 \). The only cases in which 2SLS exhibited any effective gain in prediction risk was for the smallest sample size and for well-behaved data-sampling processes, although even in these cases the gain was far from dramatic. Of course, in terms of this performance metric, the least squares estimator dominates over all sample sizes and scenarios.

4.2 Inference

We now focus on presenting empirical evidence regarding the sizes of tests of moment information validity and we also present results relating to coverage and power relating to tests of the values of structural parameters.

4.2.1 Testing Moment Equation Validity

Four different tests of moment information validity are presented in Fig. 4 and all four of the tests share the same asymptotic Chi square (\( m - k, 0 \))
distribution under $H_0$. These tests include two tests based on the usual GMM-motivated Wald-type quadratic form and use either the 2SLS or EL estimates (the W2SLS and WEL tests, respectively), as

$$W = \left( \sum_{i=1}^{n} h(Y_i, X_i, Z_i, \hat{\beta}) \right)^{-1}\left( \sum_{i=1}^{n} h(Y_i, X_i, Z_i, \hat{\beta}) h(Y_i, X_i, Z_i, \hat{\beta})' \right)$$

$$\times \left( \sum_{i=1}^{n} h(Y_i, X_i, Z_i, \hat{\beta}) \right)'$$

(12)

where $h(y_i, x_i, z_i, \hat{\beta}) \equiv z_i(y_i - x_i \hat{\beta})$, and $\hat{\beta}$ is either the 2SLS or EL estimator of $\beta$.

The empirical likelihood ratio test (ELR) is based on the ratio of constrained (by the moment equations) and unconstrained empirical likelihood functions, as defined in Section 2.2, and is calculated as

$$ELR = -2\left( \sum_{i=1}^{n} \ln(\hat{\omega}_i) + n \ln(n) \right)$$

(13)

The Lagrange multiplier test (the LM test) is based on the asymptotic normal distribution of the Lagrange multipliers in the EL optimization problem, includes a robust estimator of the asymptotic covariance matrix of this distribution (see Imbens et al., 1998), and is defined by

$$LM = n \hat{\alpha}'\hat{cov}(\hat{\alpha})^{-1} \hat{\alpha}$$

(14)

where $\hat{cov}(\hat{\alpha}) = RQ^{-1}R'$,

$$R \equiv \sum_{i=1}^{n} \hat{\omega}_i h(Y_i, X_i, Z_i, \hat{\beta}_{EL}) h(Y_i, X_i, Z_i, \hat{\beta}_{EL})'$$

(15)

and

$$Q \equiv \sum_{i=1}^{n} \hat{\omega}_i^2 h(Y_i, X_i, Z_i, \hat{\beta}_{EL}) h(Y_i, X_i, Z_i, \hat{\beta}_{EL})'$$

(16)

The target size for all of these tests in the MC experiments was 0.01.

For the well-behaved data-sampling scenarios 1–4, all of the test sizes are notably different from the target test size when $n = 25$ (see Table A.4). At $n = 50$ (see Fig. 4 and Table A.4), the WEL test has reasonably accurate size while the remaining tests still exhibit notable departures from target test size in almost all cases. Once the sample sizes is $n = 100$ or larger, all of the tests exhibit test sizes in the neighborhood of the target 0.01 size, with 2SLS tending to be the least accurate on average.
Regarding the tests of moment equation validity for the ill-behaved sampling scenarios, the performance of the 2SLS Wald-based approach is seen to be relatively erratic across the data-sampling processes. However, the majority of cases are such that the realized test size is within 0.02 of the target size of 0.01 (see Fig. 4 and Table A.4). When $n \geq 50$, all three test procedures based on the EL estimator perform reasonably well in terms of approximating the target test size of 0.01 in all but one scenario, and the results are substantially more stable than for the 2SLS estimator.

4.2.2 Coverage and Power

Empirical confidence interval coverage for $\beta_2$ is presented in Fig. 5 for $n = 50$ and $n = 100$, and results for all sample sizes are presented in Table A.5. The confidence intervals were calculated in the usual way by centering the intervals at either the 2SLS or EL estimate, and then forming interval bounds at $\pm t_{1-\tau/2}$ standard deviations from the estimate. The value $t_{1-\tau/2}$ is the $100(\tau/2)$% quantile of the $T$-distribution, with degrees of freedom equal to $n-k$, and the standard deviation used in the definition of the confidence interval is estimated based on the asymptotic covariance matrix of either the 2SLS or EL estimators. This corresponds via duality to a test of the hypothesis $H_o: \beta_2 = 2$ (truth).

The target coverage is 0.99 in these experiments, and for the well-behaved data sampling scenarios 1–4, both 2SLS and EL provide reasonably accurate coverage when $n \geq 100$. On average, EL confidence
intervals provide the more accurate coverage, with the two approaches becoming effectively indistinguishable once a sample size of \( n = 250 \) is achieved. In terms of power relative to the incorrect hypothesis \( H_a : \beta_2 = 1 \), the confidence intervals based on 2SLS are on average more powerful than the EL confidence intervals in excluding the false \( \beta_2 \) value, although again once \( n = 250 \) is achieved, the two procedures produce results that are virtually the same. Confidence interval width is, of course, one-to-one with the variance characteristics exhibited in Fig. 3 and Table A.3, and thus in all but one of the 16 well-behaved DSPs, the 2SLS procedure produces confidence intervals with slightly smaller width.

For the poorly behaved sampling scenarios 5–10 (see Fig. 5 and Table A.5) neither 2SLS nor EL provide reasonably accurate coverage for any level of \( n \) except perhaps in two cases when \( n = 250 \). Unlike the well-behaved DSPs, EL and 2SLS confidence intervals vacillate in terms of coverage superiority for all sample sizes, although the difference in performance between the two procedures diminishes as \( n \) increases. In terms of power relative to the incorrect hypothesis \( H_a : \beta_2 = 1 \), the confidence intervals based on 2SLS are in more instances more powerful than the EL confidence intervals, although the EL-based procedure is superior in one-third of the ill-behaved scenarios, and differences in power between the two procedures dissipates as \( n \) increases. Again noting that the width of confidence intervals is one-to-one with the variance characteristics exhibited in Fig. 3 and Table A.3, confidence interval width produced by the 2SLS procedure is inferior to confidence intervals generated by the EL procedure in all but two of the 24 ill-behaved DSPs.

5 OVERVIEW

In statistical models consisting of linear structural equations, the 2SLS estimator has long been the estimator of choice when the number of moment conditions–IV variables exceeded the number of unknown response parameters in the equation in question. The 2SLS estimator solves the problem of overidentification by taking a particular rank–\( k \) linear combination of the instruments. In contrast, the nontraditional EL estimator transforms the overdetermined moments problem into a set of equations that is solvable for the model parameters by imposing a functional dependence on the moment equations through choice of sample observation weights. Although both of these estimators perform well in terms of first-order asymptotics, with both procedures leading to estimators that are asymptotically efficient for estimating the response parameters of the model, questions persist as to their small sample bias and variance performance in estimation, and their coverage, interval width, and power characteristics in terms of inference.
Furthermore, in line with sampling processes often found in practice, there are questions concerning the possible estimation and inference impacts of IVs that are only weakly correlated with right-hand side endogenous variables so that response parameters are only weakly identified or determined.

Given these questions and corresponding conjectures that appear in the literature, in this chapter we have attempted to provide some empirical evidence concerning 2SLS and EL estimator performance by simulating a range of sampling processes and observing empirical repeated sampling behavior of the estimation procedures. While MC sampling results are never definitive, we feel that the base results presented in this chapter provide important insights into the relative sampling performance of these two general moment-based estimators for a range of DSPs. From this study the following finite sample performance results are suggested:

1. In well-behaved data-sampling situations both the 2SLS and EL estimators perform much like their parametric counterparts and, in terms of both estimation and inference, provide a good basis for avoiding likelihood specifications.
2. In weak instrument situations both 2SLS and EL estimators lead to sampling results exhibiting large biases and variances for estimated model parameters.
3. In weak instrument situations neither estimator performs well in terms of coverage or size relating to hypotheses about the values of model parameters.
4. When the sampling process is not well behaved and the IVs exhibit ill-posed sampling characteristics often found in practice, neither estimator dominates performance wise. In these cases the EL estimator exhibits frequent empirical superiority in many of the performance categories such as estimator precision, and parameter and prediction SEL.
5. The EL estimator faces computational problems in the weakly determined sampling designs and in some extreme cases it is difficult to obtain convergence.
6. Seemingly well-determined and viable IV sampling processes do not necessarily lead to satisfactory estimator and inference performance in small samples, as illustrated by data-sampling design 10.
7. Estimator variability and inference performance is notably effected by moderate to high correlation between the RHS \( (y, z) \) variables in the structural equation.
8. In applied work there is usually some significant level of correlation between the estimated \( y_2 \) variable(s) and the error process. This results in inflated estimation biases and the bias for
both estimators is in the directions of the LS estimator. This is notably the case for scenario 10.

9. The small sample performance of the EL estimator is encouraging and raises the question as to the possible normative performance of other EL-like (measures of divergence) estimators.

10. The strong showing of the LS estimator in several performance categories should perhaps reinvigorate thinking not only about how to combine moment conditions but also how best to combine estimators.

Overall, the results in this chapter underscore that the asymptotic optimality of the 2SLS estimator does not in any way guarantee performance superiority in sample sizes often found in empirical econometric practice. Furthermore, new nontraditional methods of estimation, such as the EL method, represent credible alternatives that deserve further attention from both applied and theoretical econometricians.

REFERENCES


Performance of Empirical Likelihood Estimator


Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54, 426–482.


**APPENDIX**

**Table A.1** Empirical Parameter and Prediction SEL

<table>
<thead>
<tr>
<th>N</th>
<th>MC no.</th>
<th>Empirical parameter SEL</th>
<th>Empirical prediction SEL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2SLS</td>
<td>EL</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0.623</td>
<td>0.797</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0.877</td>
<td>1.004</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>1.469</td>
<td>1.563</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>2.233</td>
<td>2.067</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>3.111</td>
<td>3.028</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>3.706</td>
<td>3.621</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>3.701</td>
<td>3.683</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>3.954</td>
<td>3.911</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>5.515</td>
<td>5.598</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>7.521</td>
<td>7.688</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.294</td>
<td>0.392</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.436</td>
<td>0.536</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.647</td>
<td>0.746</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>1.029</td>
<td>1.100</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>2.298</td>
<td>2.217</td>
</tr>
<tr>
<td>50</td>
<td>6</td>
<td>3.390</td>
<td>3.144</td>
</tr>
</tbody>
</table>
### Table A.1  (Continued)

<table>
<thead>
<tr>
<th>$N$</th>
<th>MC no.</th>
<th>Empirical parameter SEL</th>
<th>Empirical prediction SEL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2SLS</td>
<td>EL</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>3.349</td>
<td>3.270</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>3.408</td>
<td>3.340</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>3.441</td>
<td>3.120</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>6.632</td>
<td>6.804</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.151</td>
<td>0.224</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.212</td>
<td>0.288</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.323</td>
<td>0.406</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.571</td>
<td>0.626</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>1.427</td>
<td>1.436</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>2.780</td>
<td>2.732</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>3.164</td>
<td>3.069</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>3.453</td>
<td>3.388</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>1.718</td>
<td>1.650</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>5.489</td>
<td>5.642</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>0.056</td>
<td>0.101</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.080</td>
<td>0.130</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.126</td>
<td>0.181</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.203</td>
<td>0.258</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.540</td>
<td>0.623</td>
</tr>
<tr>
<td>250</td>
<td>6</td>
<td>1.847</td>
<td>1.863</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>3.035</td>
<td>2.964</td>
</tr>
<tr>
<td>250</td>
<td>8</td>
<td>3.193</td>
<td>3.166</td>
</tr>
<tr>
<td>250</td>
<td>9</td>
<td>0.735</td>
<td>0.762</td>
</tr>
<tr>
<td>250</td>
<td>10</td>
<td>2.809</td>
<td>3.104</td>
</tr>
</tbody>
</table>

### Table A.2  Empirical Biases

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2SLS</td>
<td>EL</td>
<td>OLS</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0.172</td>
<td>0.128</td>
<td>1.506</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0.304</td>
<td>0.246</td>
<td>1.635</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>0.462</td>
<td>0.457</td>
<td>1.787</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>0.761</td>
<td>0.727</td>
<td>2.001</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>1.246</td>
<td>1.231</td>
<td>1.519</td>
</tr>
<tr>
<td>N</td>
<td>MC no.</td>
<td>2SLS</td>
<td>EL</td>
<td>OLS</td>
</tr>
<tr>
<td>----</td>
<td>--------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>1.459</td>
<td>1.446</td>
<td>1.527</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>1.492</td>
<td>1.475</td>
<td>1.499</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>1.496</td>
<td>1.478</td>
<td>1.504</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>1.698</td>
<td>1.602</td>
<td>2.347</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>2.185</td>
<td>2.167</td>
<td>2.350</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.080</td>
<td>0.112</td>
<td>1.514</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.121</td>
<td>0.167</td>
<td>1.654</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.195</td>
<td>0.261</td>
<td>1.787</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.378</td>
<td>0.437</td>
<td>1.998</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.963</td>
<td>0.987</td>
<td>1.515</td>
</tr>
<tr>
<td>50</td>
<td>6</td>
<td>1.379</td>
<td>1.372</td>
<td>1.506</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>1.448</td>
<td>1.442</td>
<td>1.493</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>1.464</td>
<td>1.459</td>
<td>1.506</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>1.217</td>
<td>1.167</td>
<td>2.338</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>2.023</td>
<td>2.032</td>
<td>2.342</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.033</td>
<td>0.111</td>
<td>1.518</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.070</td>
<td>0.143</td>
<td>1.651</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.099</td>
<td>0.203</td>
<td>1.785</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.170</td>
<td>0.264</td>
<td>1.998</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.646</td>
<td>0.711</td>
<td>1.513</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>1.223</td>
<td>1.248</td>
<td>1.510</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>1.426</td>
<td>1.425</td>
<td>1.503</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>1.510</td>
<td>1.509</td>
<td>1.498</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>0.669</td>
<td>0.689</td>
<td>2.343</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>1.779</td>
<td>1.799</td>
<td>2.341</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>0.015</td>
<td>0.092</td>
<td>1.517</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.026</td>
<td>0.113</td>
<td>1.647</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.032</td>
<td>0.137</td>
<td>1.784</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.071</td>
<td>0.171</td>
<td>1.996</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.284</td>
<td>0.376</td>
<td>1.515</td>
</tr>
<tr>
<td>250</td>
<td>6</td>
<td>0.931</td>
<td>0.977</td>
<td>1.513</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>1.437</td>
<td>1.439</td>
<td>1.507</td>
</tr>
<tr>
<td>250</td>
<td>8</td>
<td>1.500</td>
<td>1.501</td>
<td>1.502</td>
</tr>
<tr>
<td>250</td>
<td>9</td>
<td>0.246</td>
<td>0.297</td>
<td>2.344</td>
</tr>
<tr>
<td>250</td>
<td>10</td>
<td>1.226</td>
<td>1.278</td>
<td>2.340</td>
</tr>
<tr>
<td>N</td>
<td>MC no.</td>
<td>2SLS</td>
<td>EL</td>
<td>OLS</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0.417</td>
<td>0.474</td>
<td>0.080</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0.574</td>
<td>0.601</td>
<td>0.074</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>0.923</td>
<td>0.926</td>
<td>0.072</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>1.130</td>
<td>1.037</td>
<td>0.068</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>1.383</td>
<td>1.239</td>
<td>0.080</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>1.409</td>
<td>1.305</td>
<td>0.079</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>1.331</td>
<td>1.288</td>
<td>0.079</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>1.564</td>
<td>1.506</td>
<td>0.079</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>1.123</td>
<td>1.159</td>
<td>0.034</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>0.737</td>
<td>0.780</td>
<td>0.034</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.737</td>
<td>0.780</td>
<td>0.034</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.309</td>
<td>0.317</td>
<td>0.035</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.444</td>
<td>0.452</td>
<td>0.032</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.629</td>
<td>0.632</td>
<td>0.029</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>1.270</td>
<td>1.086</td>
<td>0.038</td>
</tr>
<tr>
<td>50</td>
<td>6</td>
<td>1.184</td>
<td>1.089</td>
<td>0.037</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>1.198</td>
<td>1.112</td>
<td>0.036</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>1.028</td>
<td>0.976</td>
<td>0.016</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>0.771</td>
<td>0.718</td>
<td>0.016</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.107</td>
<td>0.126</td>
<td>0.018</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.153</td>
<td>0.168</td>
<td>0.017</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.232</td>
<td>0.244</td>
<td>0.017</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.396</td>
<td>0.386</td>
<td>0.014</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.963</td>
<td>0.846</td>
<td>0.018</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>1.228</td>
<td>1.112</td>
<td>0.018</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>1.091</td>
<td>0.984</td>
<td>0.018</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>1.143</td>
<td>1.069</td>
<td>0.018</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>0.801</td>
<td>0.768</td>
<td>0.008</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.871</td>
<td>0.764</td>
<td>0.008</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>0.039</td>
<td>0.054</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.058</td>
<td>0.069</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.094</td>
<td>0.104</td>
<td>0.006</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.144</td>
<td>0.153</td>
<td>0.005</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.439</td>
<td>0.438</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>6</td>
<td>0.950</td>
<td>0.872</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>0.951</td>
<td>0.859</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>8</td>
<td>0.931</td>
<td>0.894</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>9</td>
<td>0.470</td>
<td>0.470</td>
<td>0.003</td>
</tr>
<tr>
<td>250</td>
<td>10</td>
<td>0.559</td>
<td>0.537</td>
<td>0.003</td>
</tr>
</tbody>
</table>
### TABLE A.4  Empirical Test Sizes for Nominal Size 0.01 Chi-Square Moment Validity Test

<table>
<thead>
<tr>
<th>N</th>
<th>MC no.</th>
<th>W2SLS</th>
<th>WEL</th>
<th>ELR</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1</td>
<td>0.015</td>
<td>0.003</td>
<td>0.054</td>
<td>0.121</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0.022</td>
<td>0.005</td>
<td>0.072</td>
<td>0.134</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>0.026</td>
<td>0.002</td>
<td>0.072</td>
<td>0.131</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>0.037</td>
<td>0.008</td>
<td>0.086</td>
<td>0.142</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>0.008</td>
<td>0.002</td>
<td>0.026</td>
<td>0.062</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>0.004</td>
<td>0.000</td>
<td>0.016</td>
<td>0.043</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>0.004</td>
<td>0.001</td>
<td>0.016</td>
<td>0.044</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>0.004</td>
<td>0.001</td>
<td>0.013</td>
<td>0.036</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>0.046</td>
<td>0.016</td>
<td>0.101</td>
<td>0.160</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>0.008</td>
<td>0.001</td>
<td>0.029</td>
<td>0.064</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.013</td>
<td>0.006</td>
<td>0.031</td>
<td>0.034</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.022</td>
<td>0.007</td>
<td>0.035</td>
<td>0.034</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.023</td>
<td>0.005</td>
<td>0.026</td>
<td>0.028</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.052</td>
<td>0.014</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.131</td>
<td>0.003</td>
<td>0.021</td>
<td>0.010</td>
</tr>
<tr>
<td>50</td>
<td>6</td>
<td>0.008</td>
<td>0.001</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>0.003</td>
<td>0.000</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>0.005</td>
<td>0.001</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>0.122</td>
<td>0.020</td>
<td>0.065</td>
<td>0.057</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>0.028</td>
<td>0.003</td>
<td>0.018</td>
<td>0.021</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.016</td>
<td>0.009</td>
<td>0.021</td>
<td>0.014</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.017</td>
<td>0.008</td>
<td>0.018</td>
<td>0.012</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.027</td>
<td>0.013</td>
<td>0.024</td>
<td>0.016</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.025</td>
<td>0.007</td>
<td>0.005</td>
<td>0.010</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.025</td>
<td>0.005</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>0.008</td>
<td>0.000</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>0.006</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>0.003</td>
<td>0.000</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>0.108</td>
<td>0.016</td>
<td>0.032</td>
<td>0.018</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.071</td>
<td>0.010</td>
<td>0.017</td>
<td>0.010</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>0.011</td>
<td>0.009</td>
<td>0.012</td>
<td>0.005</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.018</td>
<td>0.013</td>
<td>0.018</td>
<td>0.007</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.019</td>
<td>0.012</td>
<td>0.016</td>
<td>0.008</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.014</td>
<td>0.009</td>
<td>0.012</td>
<td>0.005</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.030</td>
<td>0.010</td>
<td>0.014</td>
<td>0.008</td>
</tr>
<tr>
<td>250</td>
<td>6</td>
<td>0.014</td>
<td>0.006</td>
<td>0.006</td>
<td>0.003</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>0.010</td>
<td>0.004</td>
<td>0.005</td>
<td>0.001</td>
</tr>
</tbody>
</table>

(continued)
Table A.4 (Continued)

<table>
<thead>
<tr>
<th>N</th>
<th>MC no.</th>
<th>W2SLS</th>
<th>WEL</th>
<th>ELR</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>8</td>
<td>0.006</td>
<td>0.001</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>250</td>
<td>9</td>
<td>0.060</td>
<td>0.016</td>
<td>0.021</td>
<td>0.010</td>
</tr>
<tr>
<td>250</td>
<td>10</td>
<td>0.150</td>
<td>0.023</td>
<td>0.031</td>
<td>0.014</td>
</tr>
</tbody>
</table>

W2SLS and WEL are Wald-type tests, ELR is based on the empirical likelihood ratio, and LM utilizes the Lagrange multipliers of the EL-constrained optimization problem.

Table A.5 Confidence Interval Coverage (Target = 0.99) for Containing True $\beta_2 = 2$ and Power in Excluding Incorrect $\beta_2 = 1$

<table>
<thead>
<tr>
<th>N</th>
<th>MC no.</th>
<th>Coverage 2SLS</th>
<th>Coverage EL</th>
<th>Power 2SLS</th>
<th>Power EL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0.934</td>
<td>0.927</td>
<td>0.446</td>
<td>0.387</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0.902</td>
<td>0.899</td>
<td>0.444</td>
<td>0.353</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>0.855</td>
<td>0.866</td>
<td>0.456</td>
<td>0.350</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>0.754</td>
<td>0.800</td>
<td>0.524</td>
<td>0.384</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>0.823</td>
<td>0.747</td>
<td>0.497</td>
<td>0.472</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>0.798</td>
<td>0.712</td>
<td>0.502</td>
<td>0.511</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>0.818</td>
<td>0.717</td>
<td>0.478</td>
<td>0.501</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>0.820</td>
<td>0.731</td>
<td>0.482</td>
<td>0.500</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>0.472</td>
<td>0.610</td>
<td>0.706</td>
<td>0.522</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>0.361</td>
<td>0.461</td>
<td>0.788</td>
<td>0.669</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.956</td>
<td>0.960</td>
<td>0.525</td>
<td>0.456</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.934</td>
<td>0.942</td>
<td>0.486</td>
<td>0.393</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.917</td>
<td>0.935</td>
<td>0.429</td>
<td>0.327</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.869</td>
<td>0.914</td>
<td>0.452</td>
<td>0.323</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.841</td>
<td>0.811</td>
<td>0.480</td>
<td>0.404</td>
</tr>
<tr>
<td>50</td>
<td>6</td>
<td>0.821</td>
<td>0.780</td>
<td>0.498</td>
<td>0.470</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>0.810</td>
<td>0.766</td>
<td>0.509</td>
<td>0.489</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>0.804</td>
<td>0.754</td>
<td>0.514</td>
<td>0.506</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>0.608</td>
<td>0.780</td>
<td>0.620</td>
<td>0.388</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>0.377</td>
<td>0.530</td>
<td>0.784</td>
<td>0.608</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.968</td>
<td>0.976</td>
<td>0.723</td>
<td>0.666</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>0.966</td>
<td>0.970</td>
<td>0.595</td>
<td>0.525</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>0.948</td>
<td>0.968</td>
<td>0.494</td>
<td>0.429</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>0.925</td>
<td>0.947</td>
<td>0.460</td>
<td>0.398</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>$N$</th>
<th>MC no.</th>
<th>Coverage 2SLS</th>
<th>Coverage EL</th>
<th>Power 2SLS</th>
<th>Power EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>0.853</td>
<td>0.861</td>
<td>0.479</td>
<td>0.402</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>0.826</td>
<td>0.795</td>
<td>0.484</td>
<td>0.471</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>0.802</td>
<td>0.775</td>
<td>0.521</td>
<td>0.530</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>0.800</td>
<td>0.767</td>
<td>0.528</td>
<td>0.541</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>0.754</td>
<td>0.878</td>
<td>0.509</td>
<td>0.344</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.452</td>
<td>0.593</td>
<td>0.754</td>
<td>0.600</td>
</tr>
<tr>
<td>250</td>
<td>1</td>
<td>0.979</td>
<td>0.982</td>
<td>0.951</td>
<td>0.938</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
<td>0.977</td>
<td>0.978</td>
<td>0.862</td>
<td>0.840</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0.966</td>
<td>0.971</td>
<td>0.746</td>
<td>0.730</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>0.964</td>
<td>0.967</td>
<td>0.599</td>
<td>0.595</td>
</tr>
<tr>
<td>250</td>
<td>5</td>
<td>0.923</td>
<td>0.930</td>
<td>0.452</td>
<td>0.404</td>
</tr>
<tr>
<td>250</td>
<td>6</td>
<td>0.841</td>
<td>0.810</td>
<td>0.498</td>
<td>0.518</td>
</tr>
<tr>
<td>250</td>
<td>7</td>
<td>0.799</td>
<td>0.756</td>
<td>0.553</td>
<td>0.584</td>
</tr>
<tr>
<td>250</td>
<td>8</td>
<td>0.784</td>
<td>0.755</td>
<td>0.548</td>
<td>0.584</td>
</tr>
<tr>
<td>250</td>
<td>9</td>
<td>0.896</td>
<td>0.926</td>
<td>0.462</td>
<td>0.421</td>
</tr>
<tr>
<td>250</td>
<td>10</td>
<td>0.579</td>
<td>0.690</td>
<td>0.686</td>
<td>0.635</td>
</tr>
</tbody>
</table>
7

Testing for Unit Roots in Semiannual Data

Sandra G. Feltham
Ministry of Human Resources, Government of British Columbia, Victoria, British Columbia, Canada

David E. A. Giles
University of Victoria, Victoria, British Columbia, Canada

1 INTRODUCTION

There is now a well-established literature relating to the problem of testing for nonstationarity in seasonal economic time-series data. For example, some of the earlier literature on this topic can be found in Hylleberg (1992), and a more recent overview is given by Franses (1997). This issue is of considerable importance as the distinction needs to be drawn between unit roots at the zero frequency, and unit roots at some or all of the seasonal frequencies in the case of nonannual data. An incorrect identification of the nature of such unit roots would lead to inappropriate filtering of the series prior to its use in regression analysis, say, as well as inadequate testing for possible cointegration between one such series and another.

A seasonal series is one that has a spectrum with distinct peaks at the seasonal frequencies, \( \theta = (2\pi j/s) \), \( j = 1, 2, \ldots, s-1 \), where \( s \) is the number of seasons in the year. The frequencies of interest with quarterly data, for example, are at 0, 1/4, 1/2, and 3/4 cycles, or \( \theta = 0, \pi/2, \pi, \) and \( 3\pi/2 \). An integrated series is one that has an infinite mass in its spectrum at some
frequency, $\theta$. In contrast, the spectrum of a stationary series is finite, but nonzero, at all frequencies. If a seasonal series has a unit root at all of its frequencies, it is said to be “seasonally integrated”, or SI(s), and it needs to be $s$-differenced to make it stationary. Early contributions (e.g., Dickey et al., 1984) to the problem of testing for nonstationarity in seasonal time series considered only a null of SI(s), and an alternative hypothesis of stationarity. That is, they did not consider the possibility of unit roots at only the seasonal (nonzero) frequencies, and neither did they allow for unit roots at a subset of the seasonal frequencies. The more recent test proposed by Kunst (1997) also falls into this category. In the case of quarterly data, this shortcoming was rectified partially by Osborne et al. (1988), and more fully with the tests proposed by Hylleberg et al. (hereafter HEGY) (1990) and Ghysels et al. (1994). Franses (1991) and Beaulieu and Miron (1993) provided a similar testing framework for monthly data.*

Economic time-series data recorded only on a semiannual basis are quite common. For example, many companies report their financial statements on this basis, at least on a provisional basis. The purpose of this chapter is to investigate the properties of a HEGY testing framework in the case of semiannual data, and to compare it with conventional Dickey–Fuller (DF) (1981) testing for unit roots at the zero frequency in such time series. In Section 2 we describe the testing procedure, and Section 3 discusses the asymptotic distributions of our test statistics. Section 4 deals with the finite-sample percentiles for these tests, and reports some small-sample power results. The application of the semiannual HEGY tests is illustrated in Section 5, and some concluding remarks and recommendations appear in Section 6.

2 A TESTING FRAMEWORK

2.1 Background

In this section the framework for applying HEGY-type unit root tests is outlined for the case of semiannual data. Recall the autoregressive model for a stochastic seasonal process, $\chi_t$:

$$\Psi(B)x_t = \varepsilon_t,$$

where $\varepsilon_t \sim \text{iid}(0, \sigma^2)$, $\Psi(B)$ is a polynomial in the backshift operator, $B$, and the roots of $\Psi(B) = 0$ determine whether or not the series is stationary. For the purposes of the following derivation the data-generating

*See also Franses and Hobijn (1997). Smith and Taylor (1998) provide a recent elegant discussion of the HEGY tests for quarterly data.
process of $x_t$ is assumed to be free of any deterministic components. Following HEGY (1990) and Beaulieu and Miron (1993), $\Psi(B)$ can be expressed in terms of elementary polynomials and a remainder:

$$\Psi(B) = \sum_{k=1}^{2} \lambda_k \Delta(B) \frac{1 - \delta_k(B)}{\delta_k(B)} + \Delta(B)\Psi^*(B)$$

(2)

where $\delta_k(B) = 1 - (1/\theta_k)B$, $k = 1, 2$; $\Delta(B) = \prod_{k=1}^{2} \delta_k(B)$; $\lambda_k = \Psi(\theta_k) / \prod_{k=1}^{2} \delta_\theta_k$; and $\Psi^*(B)$ is a remainder with roots outside the unit circle.

The $\theta_k$’s are the unit roots. In the case of semiannual data we have $\theta_1 = 1$ and $\theta_2 = -1$, and so $\delta_1(B) = (1 - B)$, $\delta_2(B) = (1 + B)$, and $\Delta(B) = (1 - B^2)$. Substituting these into Eq. (2) yields

$$\Psi(B) = \lambda_1(1 + B) + \lambda_2(-B)(1 - B) + (1 - B^2)\Psi^*(B)$$

(3)

Next, let $\pi_1 = -\lambda_1$ and $\pi_2 = -\lambda_2$, and substitute the right-hand side of Eq. (3) into the autoregression equation $\Psi(B)x_t = \epsilon_t$, so we have

$$-\pi_1(B)(1 + B)x_t - \pi_2(-B)(1 - B)x_t + (1 - B^2)\Psi^*(B)x_t = \epsilon_t$$

This can be rewritten in the form of a regression equation on which the tests can be based:

$$\Psi^*(B)z_{3t} = \pi_1 z_{1t-1} + \pi_2 z_{2t-1} + \epsilon_t$$

(4)

where

$$z_{1t} = (1 + B)x_t = x_t + x_{t-1}$$

$$z_{2t} = -(1 - B)x_t = -(x_t - x_{t-1})$$

$$z_{3t} = (1 - B^2)x_t = x_t - x_{t-2}$$

(5)

$t = 1, 2, 3, \ldots, n$

### 2.2 Implementation of Tests

To apply the unit root tests with semiannual data, Eq. (4) is estimated by ordinary least squares (OLS). Using a “t-test,” the null hypothesis of a unit root at the zero frequency ($\pi_1 = 0$) is tested against the one-sided alternative hypothesis $\pi < 0$, which is equivalent to testing $\Psi(1) = 0$ versus the alternative of stationarity, $\Psi(1) > 0$. Similarly, the hypothesis of a unit root at the $\pi$ frequency ($\pi_2 = 0$) is tested against the alternative $\pi_2 < 0$. In addition, the “F-statistic” for $\pi_1 = \pi_2 = 0$ may be used to test if unit roots exist at both frequencies simultaneously; i.e., if the series is “seasonally integrated.” The asymptotic and finite-sample percentiles of these nonstandard “t” and “F” statistics are considered below, and various critical values are given in Table 2.
These tests are evaluated under the assumption that \( \Psi^*(B) = 1 \). However, as Beaulieu and Miron (1993) point out, if \( \Psi(B) \) is of order greater than \( S \), then \( \Psi^*(B) \neq 1 \) and the fitted model must be “augmented” with lagged values of the dependent variable as extra regressors in order to whiten the errors. This follows from the findings of Said and Dickey (1984) in the context of the familiar (augmented) Dickey–Fuller (ADF) tests for unit roots at the zero frequency. These augmentations will introduce a finite-sample distortions in the null distributions of the “t” and “F” tests that will die out asymptotically, provided that the correct number of lags is added. In particular, the number of such augmentation terms, \( q \), must be allowed to increase as the sample size increases. Beaulieu and Miron (1990) investigated these finite-sample distortions in the case of monthly data, and found that the true sizes (significance levels) of the individual unit root “t” tests were no higher than those implied by the asymptotic critical values. They therefore suggest that this distortion may cause underrejection of the unit root hypotheses in small samples with such data.

3 ASYMPTOTIC NULL DISTRIBUTIONS

3.1 HEGY-Type Tests for Semiannual Data

This section presents the asymptotic null distributions for the half-yearly data tests under consideration in this paper. Fuller (1976) proves that, when \( \rho = -1 \), the limiting null distributions of \( \hat{\rho} \) and the associated test statistics in a simple random walk model, a random walk with drift, and a random walk with drift and trend, are simply the mirror image of the limiting distributions when \( \rho = 1 \) (Dickey and Fuller, 1979). Moreover, HEGY (1990) show that the analysis of Chan and Wei (1988) can be used to extract the asymptotic distribution theory for the HEGY “t” tests from the results of Dickey and Fuller (1979) and Fuller (1976).

The asymptotic null distributions for our two “t” statistics follow directly from the analyses* of HEGY (1990) and of Beaulieu and Miron (1993). The latter provide the derivation of the asymptotic distributions under the assumption that \( \varepsilon_t \) is a martingale difference sequence with constant variance. A sequence of random scalars, \( \{\varepsilon_t\}_{t=1}^{\infty} \), is a martingale difference sequence if \( E(\varepsilon_t) = 0 \) for all \( t \) and \( E(\varepsilon_t|\varepsilon_{t-1}, \ldots, \varepsilon_1) = 0 \) for \( t = 2, 3, \ldots, n \) (Hamilton, 1994, p. 189).

*See also Phillips (1987) and Stock (1988).

The assumption of a martingale difference sequence is stronger than that of serially uncorrelated errors, but weaker than independence (Hamilton, 1994, p.190).
When the unit root tests are applied in the context of a regression model without any deterministic terms the statistic $t_1$ has the same limiting null distribution as the familiar DF “t-test” statistic:

$$
\frac{t_1}{\left(\frac{1}{\int_0^1 [W(r)]^2dr} + \frac{1}{\int_0^1 [W(r)]^2dr}\right)^{1/2}} = \frac{1}{\ln(10)} \frac{|W(1)|^2 - 1}{\ln(10)}
$$

where $W(r)$ denotes a standard Brownian motion and $\int_0^1$ is a stochastic integral (Ghysels et al., 1994; Hamilton, 1994). Further, the asymptotic null distribution of $t_2$ is the mirror image of the usual DF distribution; i.e., the negative of $t_2$ has asymptotic distribution in Eq. (6). Essentially, these results follow from the fact that $z_{1t}$ and $z_{2t}$ in Eq. (5) are asymptotically uncorrelated.

Ghysels et al. (1994) show that this relationship with the usual DF “t-tests” also holds in finite samples. By the same arguments, this will also be true for semiannual data. Further, from the results of Engle et al., (1993), they prove that, for quarterly seasonal data the asymptotic null distributions for “F” statistics for testing either (1) that the series is seasonally integrated or (2) that there are unit roots at all seasonal frequencies, are the same as the limiting distributions of the sum of the appropriate squared “t-statistics.” Similarly, the asymptotic null distribution of our $F_{12}$ statistic, for seasonally integrated half-yearly data, follows directly from the limiting distributions of the sum of the squares of $t_1$ and $t_2$. So, in the case of no drift, no trend, and no seasonal dummy variable in the fitted regression we have

$$
F_{12} \xrightarrow{L} 1 + \frac{1}{2} \left\{ \frac{\left(\int_0^1 [W_1(r)]^2dr\right)^2}{\left(\int_0^1 [W_1(r)]^2dr\right)^{1/2}} + \frac{\left(\int_0^1 [W_2(r)]^2dr\right)^2}{\left(\int_0^1 [W_2(r)]^2dr\right)^{1/2}} \right\}
$$

where $W_i(r)$ for $i = 1, 2$ denote independent standard Brownian motions (Ghysels et al., 1994).

Next, consider the asymptotic null distributions of the tests when deterministic terms are added to the fitted regression equation [Eq. (4)]. The general effects of the various deterministic terms are discussed and then the actual distributions are presented following the notation of Beaulieu and Miron (1993). The addition of drift or trend terms affects the asymptotic null distribution of $t_1$ because these components have their spectral mass at the zero frequency (HEGY, 1990). However, the asymptotic distribution of $t_2$ is independent of the drift and trend terms as $z_{2t}$ is asymptotically orthogonal to terms that are not periodic (Beaulieu and Miron, 1993). Recall that the asymptotic distribution of the $F_{12}$
statistic is related to those of the squared $t_1$ and squared $t_2$ statistics. Thus, the asymptotic null distribution of the $F_{12}$ statistic will also change with the addition of deterministic terms, as the distribution of $t_1$ changes when a drift and/or trend is added to the fitted regression.

In the case of the “$t$” statistics, the effect of adding a seasonal dummy variable into the “HEGY regression” is the reverse of that of adding constant and trend components. Once a constant term is included in the regression, the addition of a seasonal dummy variable does not affect the asymptotic null distribution of $t_1$ any further. The subsequent addition of a seasonal dummy variable does, however, change the asymptotic null distribution of $t_2$, and so the asymptotic properties of $F_{12}$ also change in this case. Again, this all follows for reasons analogous to those in the case of quarterly data. Some further insights into these, and other relationships between the tests, in the case of quarterly data are given by Smith and Taylor (1998).

The forms of the various asymptotic distributions can now be summarized. Let the distributions of $t_1$ and $t_2$, when deterministic terms are excluded from the regression equation, be denoted as

$$t_1 \stackrel{L}{\rightarrow} \frac{N_1^0}{D_1^0}, \quad t_2 \stackrel{L}{\rightarrow} \frac{N_2^0}{D_2^0}$$

Following the notation of Beaulieu and Miron (1993) $N_1^x$ and $N_2^x$ represent the part of the numerator that differs from that in the above $t_1$ and $t_2$ distributions, respectively. Similarly $D_1^x$ and $D_2^x$ represent the part of the denominator that differs. The four variations of deterministic terms in the fitted regression model are represented by $x = \mu, \tau, \xi$, and $\xi\tau$, corresponding to regressions with a constant, a constant and trend, a constant and seasonal dummy, and finally a constant, trend, and seasonal dummy. Thus, the general form of the limiting null distributions for the “$t$-statistics” can be written as

$$t_1 \stackrel{L}{\rightarrow} \frac{\int_0^1 \left[W_1(r)^2 \, dW_1\right]^2 + N_1^x}{\left[\int_0^1 \left[W_1(r)^2 \, dr + D_1^x\right]^2\right]^{1/2}}$$

$$t_2 \stackrel{L}{\rightarrow} \frac{\int_0^1 \left[W_2(r)^2 \, dW_2\right]^2 + N_2^x}{\left[\int_0^1 \left[W_2(r)^2 \, dr + D_2^x\right]^2\right]^{1/2}}$$

The five possible asymptotic null distributions for each of the test statistics are summarized in Table 1. Only the parts of the numerator and denominator that differ from the no-drift, no-trend model, are given for the “$t$” statistics. This is done to provide visual representation of the
distributions that are the same. The full asymptotic null distribution of the $F_{12}$ statistic is given in terms of these components.

### 3.2 Dickey–Fuller Tests and Semiannual Data

Several properties of the DF “t-test” have been analyzed in the context of seasonal data. For example, HEGY (1990) showed it to have asymptotic equivalence to their $t_1$ statistic in the “no drift/no trend” case. Ghysels et al. (1994) demonstrated that when unit roots exist at seasonal frequencies the DF “t-test” is still a valid test for unit roots at the zero frequency, provided that the usual DF regression is appropriately “augmented” with $l \geq s - 1$ lagged values of the differenced series. They use Monte Carlo simulations to see how the existence of seasonal unit roots influences the finite-sample size and power of the test. Their results show that, in the case of quarterly data, when $l < s - 1$ the size of the DF test is severely distorted. When $l \geq s - 1$ the size is close to its nominal size, provided that the data-generating process is free of negative moving average components.

Although we have not explicitly explored here these characteristics of the DF “t-test” in the context of semiannual data, intuitively it is clear that corresponding results to those above will hold (with $s = 2$).

### 3.3 Moving-Average Errors

Although the HEGY test provides for identification of roots at different frequencies, it has been found to suffer from size distortions from negative MA components in the data-generating process. In particular, for the data-generating process:

$$Y_t = \delta_d Y_{t-d} + u_t$$

where $u_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_4 \epsilon_{t-4}$, and $\theta_1 = -0.9$, the near cancellation of $(1 - B)$ in the autoregressive component of $u_t$ makes the unit root at the zero frequency difficult to detect. This also explains why the DF test suffers from size distortions with negative MA components, because it has the same distribution as the HEGY $t_1$ test when $l \geq s - 1$. This is not a trivial problem as such situations arise frequently in practice.

A large bias in size or very low power was also found to occur when seasonal dummies are in the data-generating process, but are not included in the regression. This provides justification for including deterministic terms or additional lags even though they may be irrelevant and thus may reduce the power of the test.
<table>
<thead>
<tr>
<th>Model</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$F_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND, NT, NS</td>
<td>$\frac{N_0^0}{D_1^0}$</td>
<td>$\frac{N_2^0}{D_2^0}$</td>
<td>$N_{12}^0 = \left( \frac{N_0^0}{D_1^0} \right)^2 + \left( \frac{N_2^0}{D_2^0} \right)^2$</td>
</tr>
<tr>
<td>D, NT, NS</td>
<td>$N_1^0 = -W_1(1) \int_0^1 W_1(1) dr$</td>
<td>$N_2^0 = 0$</td>
<td>$\left( \int_0^1 W_1(r)^2 dW_1 \right)^2 + N_1^0$</td>
</tr>
<tr>
<td>D, T, NS</td>
<td>$N_1^0 = -4W_1(1) \int_0^1 W_1(r) dr$</td>
<td>$N_2^0 = 0$</td>
<td>$\left( \int_0^1 W_1(r)^2 dW_1 \right)^2 + N_1^0$</td>
</tr>
</tbody>
</table>

$D_{1}^i = -4 \left( \int_0^1 W_1(r) dr \right)^2 - 12 \left( \int_0^1 r W_1(r) dr \right)^2 + 12 \int_0^1 W_1(r) dr \int_0^1 r W_1(r) dr$
\[
\begin{align*}
N_1^t = N_1^e & \quad N_2^t = -W_2(1) \int_0^1 W_2(1) dr \\
D_1^t = D_1^e & \quad D_2^t = -\left( \int_0^1 W_2(1) dr \right)^2
\end{align*}
\]
4 FINITE-SAMPLE RESULTS

4.1 Critical Values

We have investigated the small-sample distributions of our three semiannual unit root tests under both the null and alternative hypotheses using Monte Carlo simulations. All of these simulations were implemented using SHAZAM (1997) on a DEC Alpha 3000 workstation. The data-generating process used was $x_t = x_{t-2} + \varepsilon_t$, where the $\varepsilon_t$ values were generated by the normal random number generator in SHAZAM. Five different variations of the basic HEGY auxiliary regression were fitted to the generated data: drift, seasonal dummy variable, and trend (D, S, T); drift, seasonal dummy, and no trend (D, S, NT); drift, no seasonal dummy variable, and trend (D, NS, T); drift, no seasonal dummy variable, and no trend (D, NS, NT); and no drift, no seasonal dummy variable, and no trend (ND, NS, NT).

The initial value for $y_t$ was set to zero and then the first 200 observations were dropped to “wash out” the effect of this initial value. All simulations were based on 20,000 repetitions. Critical values associated with the 1st, 5th, and 10th percentiles of the underlying distributions were generated for the “$t_1$ test” of $H_0: \pi_1 = 0$ and for the “$t_2$ test” of $H_0: \pi_2 = 0$. Corresponding critical values were calculated for the “$F_{12}$ test” of $H_0: \pi_1 = \pi_2 = 0$ by determining the 90th, 95th, and 99th percentiles of the empirical distributions. These results are presented in Table 2.

There is no discernable change in the $t_2$ critical values when a drift and/or trend are added to the fitted regressions, as was indicated by the asymptotic distribution theory in Section 3. In contrast, the critical values for $t_1$ change when either a drift or trend is added to the fitted model on which the tests are based. The subsequent addition of a seasonal dummy variable does not then change the finite sample distribution of $t_1$, but it does change that of $t_2$. As predicted by the asymptotic distribution theory, the “$F_{12}$ statistic” critical values change with each variation to the deterministic components in the fitted regression, as its distribution is the same as the distribution of the sum of the squares of $t_1$ and $t_2$.

It is also worth noting the very close accordance between the no-drift, no-trend values for $t_1$ in Table 2, the corresponding DF critical values of MacKinnon (1991), and the HEGY (1990) $\pi_1$ values. For example, with a 10% significance level and $n = 200$, the critical values are $-1.6163$ in Table 2, $-1.6165$ from MacKinnon, and $-1.62$ from HEGY. The critical values for $t_2$ (i.e., for testing $H_0: \pi_2 = 0$) in Table 2 can be compared with the HEGY (1990) critical values for the fitted model with no deterministic terms. For the 10% level and $n = 200$ the two statistics are $-1.6112$ and $-1.61$, respectively.
The results in Table 2 are based on a data-generating process (d.g.p.) of \( x_t = x_{t-2} + \varepsilon_t \). That is, unit roots are present at both the zero and \( \pi \) frequencies—the semi-annual series is “seasonally integrated.” An alternative approach is to consider a d.g.p. that has a unit root at only one of these frequencies, but still base the unit roots tests on the HEGY-type integrating regression as in Eq. (4). Critical values were also generated for these tests. In the test for a unit root only at the zero frequency, a simple random walk was used as the d.g.p.; in the second case the d.g.p. had a unit root at the \( \pi \) frequency, but not at the zero frequency. For completeness, these results appear in Table 3.

4.2 Powers of the Tests

The Monte Carlo experiment was also used to simulate the size-adjusted powers of the test statistics developed in this chapter, for the following three fitted regressions: drift, seasonal dummy variable, and trend (D, S, T); drift, seasonal dummy variable, and no trend (D, S, NT); and drift, no seasonal dummy, with trend (D, NS, T).

Using the d.g.p. of Table 2, namely, \( x_t = \rho x_{t-2} + \varepsilon_t \), the size-adjusted powers of the tests of \( \pi_1 = 0, \pi_2 = 0, \) and \( \pi_1 = \pi_2 = 0 \) were simulated for \( \rho = [0.0(0.1)1.0] \) and various sample sizes (\( n \)). Each power curve was generated using 20,000 repetitions. The results of these empirical power simulations for the \( t_1, f_2, \) and \( F_{12} \) tests are given in Tables 4–6, for the case where the d.g.p. is as in Table 2; and in Tables 7 and 8 for the case where the d.g.p. is as in Table 3. Illustrative plots appear in Figs. 1–3 and Figs. 4 and 5, respectively.

These results show that the power curves for large sample sizes have the basic expected shape. That is, they move towards 100% as \( \rho \) moves away from 1. There is reasonable power when \( n = 100 \), but very low size-adjusted power when the sample size is small. This is no surprise given other results for similar such unit root tests. The patterns in the results as the sample size is increased reflect the consistency of the tests. Changing the significance level of the test yields the usual trade off between the size and power. As expected, the tests based on auxiliary regressions without a drift or trend have more power than their counterparts with deterministic terms, because the d.g.p. does not include deterministic terms.

5 SOME APPLICATIONS

The application of the testing procedures discussed in this chapter is illustrated here for four quite different semiannual economic time series. In each case, we compare the results obtained when the possibility of seasonal...
<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>( \pi_1 = 0 )</th>
<th>( \pi_2 = 0 )</th>
<th>( \pi_1 = \pi_2 = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>10000</td>
<td></td>
<td>5000</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D, NS, NT</td>
<td>-4.0211</td>
<td>-3.4659</td>
<td>-3.1765</td>
<td>-2.4668</td>
</tr>
<tr>
<td></td>
<td>-3.4490</td>
<td>-2.8577</td>
<td>-2.5624</td>
<td>-2.5216</td>
</tr>
<tr>
<td></td>
<td>-2.7309</td>
<td>-1.9283</td>
<td>-1.5399</td>
<td>-2.7428</td>
</tr>
<tr>
<td></td>
<td>-2.6158</td>
<td>-1.9105</td>
<td>-1.5636</td>
<td>-2.6117</td>
</tr>
<tr>
<td></td>
<td>-2.6042</td>
<td>-1.9164</td>
<td>-1.5943</td>
<td>-2.5846</td>
</tr>
<tr>
<td></td>
<td>-2.5915</td>
<td>-1.9499</td>
<td>-1.6163</td>
<td>-2.5291</td>
</tr>
<tr>
<td></td>
<td>-2.6220</td>
<td>-1.9774</td>
<td>-1.6621</td>
<td>-2.4661</td>
</tr>
<tr>
<td></td>
<td>-2.6273</td>
<td>-1.9975</td>
<td>-1.6690</td>
<td>-2.5145</td>
</tr>
<tr>
<td>Model</td>
<td>N</td>
<td>1%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-4.0394</td>
<td>-3.4492</td>
<td>-3.1518</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>-4.0074</td>
<td>-3.4647</td>
<td>-3.1744</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>-4.0408</td>
<td>-3.4647</td>
<td>-3.1822</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-3.4967</td>
<td>-2.8749</td>
<td>-2.5696</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>-3.4758</td>
<td>-2.8964</td>
<td>-2.6018</td>
</tr>
<tr>
<td>D, NS, T</td>
<td>20</td>
<td>-4.5285</td>
<td>-3.6806</td>
<td>-3.2892</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-4.1618</td>
<td>-3.5155</td>
<td>-3.2006</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-4.0525</td>
<td>-3.4448</td>
<td>-3.1542</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-4.0103</td>
<td>-3.4370</td>
<td>-3.1307</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>-4.0074</td>
<td>-3.4642</td>
<td>-3.1746</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>-4.0411</td>
<td>-3.4643</td>
<td>-3.1821</td>
</tr>
<tr>
<td>D, NS, NT</td>
<td>20</td>
<td>-3.7839</td>
<td>-2.9961</td>
<td>-2.6377</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-3.5449</td>
<td>-2.9116</td>
<td>-2.6002</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-3.4896</td>
<td>-2.8807</td>
<td>-2.5719</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-3.4728</td>
<td>-2.8623</td>
<td>-2.5638</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>-3.4745</td>
<td>-2.9074</td>
<td>-2.5918</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>-3.4757</td>
<td>-2.8967</td>
<td>-2.6014</td>
</tr>
<tr>
<td>ND, NS, NT</td>
<td>20</td>
<td>-2.7678</td>
<td>-1.9597</td>
<td>-1.5652</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-2.6501</td>
<td>-1.9284</td>
<td>-1.5774</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-2.6190</td>
<td>-1.9283</td>
<td>-1.6007</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-2.6011</td>
<td>-1.9519</td>
<td>-1.6247</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>-2.6332</td>
<td>-1.9771</td>
<td>-1.6620</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>-2.6241</td>
<td>-1.9972</td>
<td>-1.6697</td>
</tr>
</tbody>
</table>
## Table 4  
Power of the $t_1$ Test; d.g.p.: $(1 - \beta^2)x_t = \varepsilon_t$

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
</table>
| D, S, T  
1% |
| $n = 20$ | 1.00 | 1.44 | 1.46 | 1.86 | 2.43 | 2.84 | 3.99 | 5.80 | 8.28 | 11.71 | 18.17 |
| $n = 50$ | 1.00 | 1.28 | 2.05 | 3.61 | 7.09 | 12.93 | 24.56 | 41.94 | 61.44 | 79.63 | 92.04 |
| $n = 100$ | 1.00 | 1.96 | 5.88 | 17.41 | 41.18 | 71.77 | 91.67 | 99.07 | 99.91 | 100.00 | 100.00 |
| 5% |
| $n = 20$ | 5.00 | 5.83 | 6.63 | 7.42 | 9.31 | 11.73 | 14.72 | 19.55 | 26.65 | 33.98 | 44.89 |
| $n = 50$ | 5.00 | 6.26 | 8.92 | 15.43 | 26.00 | 40.87 | 59.31 | 77.72 | 90.12 | 96.65 | 98.27 |
| $n = 100$ | 5.00 | 9.06 | 22.27 | 48.32 | 78.19 | 94.92 | 99.51 | 99.99 | 100.00 | 100.00 | 100.00 |
| 10% |
| $n = 20$ | 10.00 | 11.31 | 12.44 | 14.01 | 17.14 | 21.09 | 25.60 | 32.84 | 41.72 | 50.93 | 61.79 |
| $n = 50$ | 10.00 | 12.38 | 17.38 | 27.48 | 42.57 | 59.78 | 76.74 | 90.27 | 96.61 | 99.15 | 98.89 |
| $n = 100$ | 10.00 | 16.83 | 36.61 | 66.88 | 90.83 | 98.68 | 99.90 | 100.00 | 100.00 | 100.00 | 100.00 |
| D, S, NT  
1% |
| $n = 20$ | 1.00 | 1.45 | 1.69 | 2.07 | 3.23 | 4.36 | 6.38 | 10.15 | 15.33 | 21.65 | 32.44 |
| $n = 50$ | 1.00 | 1.71 | 3.32 | 7.00 | 15.20 | 28.10 | 47.50 | 68.91 | 85.55 | 94.78 | 98.79 |
| $n = 100$ | 1.00 | 3.13 | 12.40 | 35.15 | 68.80 | 91.97 | 99.03 | 99.98 | 100.00 | 100.00 | 100.00 |
| 5% |
| $n = 20$ | 5.00 | 6.52 | 7.85 | 10.09 | 14.02 | 18.75 | 24.53 | 34.09 | 44.27 | 55.76 | 68.28 |
| $n = 50$ | 5.00 | 8.13 | 14.63 | 26.70 | 44.85 | 65.44 | 82.71 | 94.22 | 98.39 | 99.61 | 99.95 |
| $n = 100$ | 5.00 | 13.15 | 37.73 | 72.53 | 94.80 | 99.41 | 99.97 | 100.00 | 100.00 | 100.00 | 100.00 |

*(continued)*
<table>
<thead>
<tr>
<th>$\rho$</th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
<td>12.45</td>
<td>15.14</td>
<td>18.54</td>
<td>24.78</td>
<td>31.69</td>
<td>40.04</td>
<td>51.44</td>
<td>62.87</td>
<td>73.09</td>
<td>82.97</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.00</td>
<td>15.39</td>
<td>25.85</td>
<td>42.71</td>
<td>63.87</td>
<td>82.32</td>
<td>93.21</td>
<td>98.36</td>
<td>99.60</td>
<td>99.93</td>
<td>99.99</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.01</td>
<td>24.78</td>
<td>58.10</td>
<td>88.19</td>
<td>98.89</td>
<td>99.91</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>D, NS, T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
<td>1.44</td>
<td>1.62</td>
<td>1.88</td>
<td>2.61</td>
<td>3.28</td>
<td>4.24</td>
<td>6.57</td>
<td>9.07</td>
<td>13.06</td>
<td>19.97</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
<td>1.34</td>
<td>2.25</td>
<td>3.87</td>
<td>7.52</td>
<td>13.80</td>
<td>26.04</td>
<td>43.96</td>
<td>63.38</td>
<td>80.99</td>
<td>92.87</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>1.93</td>
<td>5.97</td>
<td>17.46</td>
<td>41.32</td>
<td>71.93</td>
<td>91.84</td>
<td>99.12</td>
<td>99.91</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>5.01</td>
<td>6.50</td>
<td>7.53</td>
<td>8.20</td>
<td>10.29</td>
<td>13.13</td>
<td>16.31</td>
<td>21.92</td>
<td>29.46</td>
<td>37.68</td>
<td>48.91</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>5.00</td>
<td>6.57</td>
<td>9.42</td>
<td>15.81</td>
<td>26.92</td>
<td>41.86</td>
<td>60.21</td>
<td>78.69</td>
<td>90.78</td>
<td>97.00</td>
<td>98.34</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>5.00</td>
<td>8.91</td>
<td>22.19</td>
<td>48.19</td>
<td>78.16</td>
<td>94.89</td>
<td>99.50</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
<td>12.41</td>
<td>13.81</td>
<td>15.52</td>
<td>19.00</td>
<td>23.60</td>
<td>28.20</td>
<td>36.01</td>
<td>45.03</td>
<td>54.95</td>
<td>65.92</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.00</td>
<td>12.85</td>
<td>18.01</td>
<td>28.12</td>
<td>43.38</td>
<td>60.57</td>
<td>77.43</td>
<td>90.85</td>
<td>96.80</td>
<td>99.22</td>
<td>99.90</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.00</td>
<td>16.95</td>
<td>36.92</td>
<td>67.22</td>
<td>91.00</td>
<td>98.73</td>
<td>99.93</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
**Table 5**  Power of the $t_2$ Test; d.g.p.: $(1 - \beta^2)x_t = \epsilon_t$

<table>
<thead>
<tr>
<th></th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D, S, T</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
<td>1.22</td>
<td>1.62</td>
<td>2.28</td>
<td>3.15</td>
<td>4.44</td>
<td>6.81</td>
<td>10.04</td>
<td>15.06</td>
<td>22.33</td>
<td>32.01</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
<td>1.48</td>
<td>2.81</td>
<td>6.67</td>
<td>13.39</td>
<td>25.50</td>
<td>44.65</td>
<td>66.16</td>
<td>83.26</td>
<td>93.75</td>
<td>98.46</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>3.06</td>
<td>10.89</td>
<td>32.64</td>
<td>65.30</td>
<td>96.60</td>
<td>99.95</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>1%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>5.00</td>
<td>5.89</td>
<td>7.43</td>
<td>9.48</td>
<td>12.87</td>
<td>17.46</td>
<td>23.38</td>
<td>31.23</td>
<td>41.57</td>
<td>52.30</td>
<td>64.86</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>5.00</td>
<td>7.85</td>
<td>13.94</td>
<td>25.96</td>
<td>43.72</td>
<td>64.57</td>
<td>82.26</td>
<td>93.80</td>
<td>98.32</td>
<td>99.68</td>
<td>99.96</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>5.00</td>
<td>14.04</td>
<td>38.52</td>
<td>70.91</td>
<td>93.85</td>
<td>99.45</td>
<td>99.97</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D, S, NT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
<td>11.50</td>
<td>14.57</td>
<td>18.13</td>
<td>23.20</td>
<td>30.16</td>
<td>38.55</td>
<td>48.95</td>
<td>60.08</td>
<td>70.59</td>
<td>80.85</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.00</td>
<td>15.08</td>
<td>25.43</td>
<td>42.34</td>
<td>63.43</td>
<td>81.61</td>
<td>93.11</td>
<td>98.33</td>
<td>99.65</td>
<td>99.96</td>
<td>100.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.00</td>
<td>24.91</td>
<td>58.06</td>
<td>87.90</td>
<td>98.70</td>
<td>99.94</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>1%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
<td>1.20</td>
<td>1.56</td>
<td>2.08</td>
<td>2.93</td>
<td>4.12</td>
<td>6.13</td>
<td>9.39</td>
<td>14.33</td>
<td>21.43</td>
<td>31.31</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
<td>1.53</td>
<td>3.08</td>
<td>7.02</td>
<td>14.08</td>
<td>26.59</td>
<td>46.06</td>
<td>67.71</td>
<td>84.40</td>
<td>94.36</td>
<td>98.64</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>3.07</td>
<td>10.91</td>
<td>32.18</td>
<td>64.86</td>
<td>90.33</td>
<td>98.64</td>
<td>99.95</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>(continued)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 5 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>10.00</td>
<td>11.87</td>
<td>14.57</td>
<td>18.31</td>
<td>23.90</td>
<td>30.98</td>
<td>39.56</td>
<td>50.11</td>
<td>61.53</td>
<td>71.88</td>
<td>82.17</td>
</tr>
<tr>
<td>n = 50</td>
<td>9.99</td>
<td>15.15</td>
<td>25.46</td>
<td>42.58</td>
<td>63.89</td>
<td>81.94</td>
<td>93.35</td>
<td>98.42</td>
<td>99.67</td>
<td>99.97</td>
<td>100.00</td>
</tr>
<tr>
<td>n = 100</td>
<td>10.00</td>
<td>25.01</td>
<td>58.28</td>
<td>87.59</td>
<td>98.65</td>
<td>99.91</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### D, N S, T

<table>
<thead>
<tr>
<th></th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>1.00</td>
<td>2.45</td>
<td>4.14</td>
<td>6.76</td>
<td>11.06</td>
<td>17.20</td>
<td>25.07</td>
<td>36.50</td>
<td>48.87</td>
<td>62.21</td>
<td>74.77</td>
</tr>
<tr>
<td>n = 50</td>
<td>1.00</td>
<td>4.77</td>
<td>12.31</td>
<td>27.43</td>
<td>51.28</td>
<td>74.27</td>
<td>90.17</td>
<td>97.39</td>
<td>99.46</td>
<td>99.92</td>
<td>100.00</td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
<td>11.21</td>
<td>42.15</td>
<td>80.00</td>
<td>97.39</td>
<td>99.87</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 20</td>
<td>5.00</td>
<td>13.57</td>
<td>20.93</td>
<td>30.92</td>
<td>43.56</td>
<td>56.70</td>
<td>69.29</td>
<td>80.78</td>
<td>89.25</td>
<td>94.13</td>
<td>97.39</td>
</tr>
<tr>
<td>n = 50</td>
<td>5.00</td>
<td>20.52</td>
<td>44.15</td>
<td>70.51</td>
<td>88.79</td>
<td>97.65</td>
<td>99.56</td>
<td>99.95</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>n = 100</td>
<td>5.00</td>
<td>40.84</td>
<td>85.60</td>
<td>98.85</td>
<td>99.96</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### 10%  

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 20</td>
<td>10.00</td>
<td>25.46</td>
<td>37.51</td>
<td>50.40</td>
<td>65.41</td>
<td>78.03</td>
<td>86.74</td>
<td>93.30</td>
<td>96.60</td>
<td>98.50</td>
<td>99.37</td>
</tr>
<tr>
<td>n = 50</td>
<td>10.00</td>
<td>37.39</td>
<td>67.40</td>
<td>89.11</td>
<td>97.87</td>
<td>99.67</td>
<td>99.95</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>n = 100</td>
<td>10.00</td>
<td>64.16</td>
<td>96.48</td>
<td>99.88</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
TABLE 6  

<table>
<thead>
<tr>
<th></th>
<th>0.00</th>
<th>0.10</th>
<th>0.20</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
<th>0.80</th>
<th>0.90</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D, S, T )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n = 20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D, S, T )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n = 20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D, S, T )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n = 20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D, S, T )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th></th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
<td>12.16</td>
<td>15.46</td>
<td>20.72</td>
<td>27.39</td>
<td>36.39</td>
<td>47.89</td>
<td>59.82</td>
<td>72.00</td>
<td>81.90</td>
<td>89.15</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.00</td>
<td>18.35</td>
<td>36.10</td>
<td>59.53</td>
<td>81.81</td>
<td>94.72</td>
<td>99.02</td>
<td>99.81</td>
<td>99.98</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.01</td>
<td>33.09</td>
<td>78.42</td>
<td>98.36</td>
<td>99.96</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>D, NS, T</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.01</td>
<td>1.37</td>
<td>1.82</td>
<td>2.77</td>
<td>4.04</td>
<td>6.16</td>
<td>9.90</td>
<td>14.79</td>
<td>22.98</td>
<td>33.89</td>
<td>45.70</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
<td>2.27</td>
<td>5.71</td>
<td>13.97</td>
<td>30.22</td>
<td>54.30</td>
<td>78.31</td>
<td>92.33</td>
<td>98.02</td>
<td>99.71</td>
<td>99.94</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>5.33</td>
<td>26.62</td>
<td>68.92</td>
<td>94.98</td>
<td>99.72</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>5%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>5.01</td>
<td>6.63</td>
<td>8.88</td>
<td>12.03</td>
<td>16.54</td>
<td>23.05</td>
<td>31.73</td>
<td>42.99</td>
<td>55.22</td>
<td>68.18</td>
<td>78.25</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>5.01</td>
<td>10.59</td>
<td>22.10</td>
<td>42.16</td>
<td>66.56</td>
<td>86.34</td>
<td>96.44</td>
<td>99.25</td>
<td>99.91</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>5.01</td>
<td>20.11</td>
<td>61.40</td>
<td>93.68</td>
<td>99.70</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>10%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
<td>13.12</td>
<td>16.67</td>
<td>22.01</td>
<td>28.39</td>
<td>38.30</td>
<td>48.80</td>
<td>60.92</td>
<td>72.59</td>
<td>82.47</td>
<td>89.20</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.01</td>
<td>19.14</td>
<td>36.90</td>
<td>59.88</td>
<td>81.56</td>
<td>94.57</td>
<td>98.91</td>
<td>99.81</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.00</td>
<td>34.16</td>
<td>78.64</td>
<td>98.25</td>
<td>99.95</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Table 7</td>
<td>Power of the $t_1$ Test; d.g.p.: $(1 - \beta)x_t = \epsilon_t$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>D, S, T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
<td>1.13</td>
<td>1.38</td>
<td>2.03</td>
<td>2.58</td>
<td>3.82</td>
<td>5.19</td>
<td>6.49</td>
<td>8.65</td>
<td>11.15</td>
<td>13.99</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
<td>1.91</td>
<td>4.05</td>
<td>10.49</td>
<td>20.56</td>
<td>34.37</td>
<td>49.95</td>
<td>65.13</td>
<td>75.84</td>
<td>85.13</td>
<td>91.46</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>4.69</td>
<td>23.67</td>
<td>58.87</td>
<td>86.32</td>
<td>97.21</td>
<td>99.51</td>
<td>99.94</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>5.00</td>
<td>5.60</td>
<td>6.83</td>
<td>9.01</td>
<td>11.56</td>
<td>15.53</td>
<td>19.58</td>
<td>24.24</td>
<td>28.59</td>
<td>34.57</td>
<td>39.58</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>5.00</td>
<td>8.22</td>
<td>16.78</td>
<td>33.01</td>
<td>51.80</td>
<td>69.10</td>
<td>82.38</td>
<td>90.67</td>
<td>94.97</td>
<td>97.96</td>
<td>99.01</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>5.00</td>
<td>18.46</td>
<td>56.69</td>
<td>88.89</td>
<td>98.34</td>
<td>99.81</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
<td>11.10</td>
<td>13.44</td>
<td>17.14</td>
<td>21.75</td>
<td>27.02</td>
<td>33.49</td>
<td>38.91</td>
<td>44.95</td>
<td>52.21</td>
<td>57.85</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.00</td>
<td>15.48</td>
<td>29.41</td>
<td>50.08</td>
<td>69.36</td>
<td>83.11</td>
<td>92.20</td>
<td>96.58</td>
<td>98.35</td>
<td>99.47</td>
<td>99.78</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.01</td>
<td>31.40</td>
<td>75.05</td>
<td>96.06</td>
<td>99.65</td>
<td>99.97</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>D, S, NT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
<td>1.58</td>
<td>2.03</td>
<td>3.51</td>
<td>5.19</td>
<td>7.92</td>
<td>10.86</td>
<td>14.88</td>
<td>18.34</td>
<td>23.80</td>
<td>28.44</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
<td>3.16</td>
<td>8.77</td>
<td>22.26</td>
<td>41.03</td>
<td>60.13</td>
<td>76.25</td>
<td>86.96</td>
<td>93.07</td>
<td>97.08</td>
<td>98.63</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
<td>8.69</td>
<td>42.70</td>
<td>81.94</td>
<td>96.97</td>
<td>99.64</td>
<td>99.96</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>ρ</th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>5.00</td>
<td>7.39</td>
<td>10.32</td>
<td>15.72</td>
<td>21.38</td>
<td>28.45</td>
<td>36.98</td>
<td>44.13</td>
<td>50.76</td>
<td>58.72</td>
<td>66.34</td>
</tr>
<tr>
<td>n = 50</td>
<td>5.00</td>
<td>12.92</td>
<td>30.76</td>
<td>55.31</td>
<td>76.85</td>
<td>89.11</td>
<td>95.74</td>
<td>98.53</td>
<td>99.35</td>
<td>99.80</td>
<td>99.95</td>
</tr>
<tr>
<td>n = 100</td>
<td>5.00</td>
<td>31.44</td>
<td>81.52</td>
<td>98.14</td>
<td>99.91</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>10.00</td>
<td>14.12</td>
<td>19.21</td>
<td>27.02</td>
<td>35.45</td>
<td>44.44</td>
<td>53.27</td>
<td>61.87</td>
<td>68.62</td>
<td>75.14</td>
<td>80.36</td>
</tr>
<tr>
<td>n = 50</td>
<td>10.00</td>
<td>23.15</td>
<td>48.71</td>
<td>73.44</td>
<td>89.41</td>
<td>96.17</td>
<td>98.83</td>
<td>99.67</td>
<td>99.85</td>
<td>99.98</td>
<td>100.00</td>
</tr>
<tr>
<td>n = 100</td>
<td>10.00</td>
<td>49.72</td>
<td>92.97</td>
<td>99.65</td>
<td>99.97</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>D, NS, T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>1.00</td>
<td>1.11</td>
<td>1.37</td>
<td>2.08</td>
<td>2.62</td>
<td>3.71</td>
<td>5.00</td>
<td>6.73</td>
<td>8.88</td>
<td>11.49</td>
<td>14.22</td>
</tr>
<tr>
<td>n = 50</td>
<td>1.00</td>
<td>1.82</td>
<td>3.98</td>
<td>10.36</td>
<td>20.40</td>
<td>34.16</td>
<td>49.88</td>
<td>64.93</td>
<td>75.91</td>
<td>85.14</td>
<td>91.43</td>
</tr>
<tr>
<td>n = 100</td>
<td>1.00</td>
<td>4.53</td>
<td>23.14</td>
<td>58.13</td>
<td>85.80</td>
<td>97.05</td>
<td>99.48</td>
<td>99.93</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>5.00</td>
<td>5.67</td>
<td>6.88</td>
<td>8.98</td>
<td>11.72</td>
<td>15.64</td>
<td>19.60</td>
<td>24.31</td>
<td>29.05</td>
<td>34.75</td>
<td>39.66</td>
</tr>
<tr>
<td>n = 50</td>
<td>5.00</td>
<td>8.18</td>
<td>16.78</td>
<td>32.80</td>
<td>51.80</td>
<td>68.99</td>
<td>82.35</td>
<td>90.70</td>
<td>94.97</td>
<td>97.93</td>
<td>99.03</td>
</tr>
<tr>
<td>n = 100</td>
<td>5.00</td>
<td>18.69</td>
<td>57.03</td>
<td>89.08</td>
<td>98.43</td>
<td>99.80</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 20</td>
<td>10.00</td>
<td>11.15</td>
<td>13.32</td>
<td>16.91</td>
<td>21.67</td>
<td>26.73</td>
<td>33.05</td>
<td>39.03</td>
<td>45.08</td>
<td>52.21</td>
<td>58.04</td>
</tr>
<tr>
<td>n = 50</td>
<td>10.00</td>
<td>15.46</td>
<td>29.16</td>
<td>49.85</td>
<td>69.23</td>
<td>83.01</td>
<td>92.34</td>
<td>96.61</td>
<td>98.35</td>
<td>99.44</td>
<td>99.79</td>
</tr>
<tr>
<td>n = 100</td>
<td>9.99</td>
<td>31.37</td>
<td>74.83</td>
<td>96.07</td>
<td>99.63</td>
<td>99.97</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Table 8  Power of the $t_2$ Test; d.g.p.: $(1 + \beta)x_t = \varepsilon_t$

<table>
<thead>
<tr>
<th></th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>D, S, T 1%</td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
</tr>
<tr>
<td>D, S, NT 1%</td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>5.00</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>5.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>4.99</td>
</tr>
<tr>
<td>D, S, NT 5%</td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>10.00</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>10.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>10.00</td>
</tr>
<tr>
<td>D, S, NT 10%</td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>1.00</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>1.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>1.00</td>
</tr>
<tr>
<td>D, S, NT 5%</td>
<td></td>
</tr>
<tr>
<td>$n = 20$</td>
<td>4.99</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>5.00</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>4.99</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>1.00</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.60</th>
<th>0.50</th>
<th>0.40</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>n = 20</td>
<td>10.00</td>
<td>13.60</td>
<td>19.13</td>
<td>26.41</td>
<td>34.19</td>
<td>43.78</td>
<td>52.74</td>
<td>60.94</td>
<td>67.69</td>
<td>74.99</td>
<td>79.71</td>
</tr>
<tr>
<td></td>
<td>n = 50</td>
<td>10.00</td>
<td>23.42</td>
<td>48.55</td>
<td>73.96</td>
<td>88.77</td>
<td>96.17</td>
<td>98.94</td>
<td>99.54</td>
<td>99.90</td>
<td>99.99</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>n = 100</td>
<td>10.00</td>
<td>50.48</td>
<td>92.79</td>
<td>99.68</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>D, NS, T</td>
<td>1%</td>
<td>n = 20</td>
<td>1.00</td>
<td>3.88</td>
<td>8.28</td>
<td>14.50</td>
<td>22.22</td>
<td>31.02</td>
<td>41.14</td>
<td>50.60</td>
<td>59.17</td>
<td>67.26</td>
</tr>
<tr>
<td></td>
<td>n = 50</td>
<td>1.00</td>
<td>10.96</td>
<td>35.38</td>
<td>64.81</td>
<td>85.81</td>
<td>94.66</td>
<td>98.47</td>
<td>99.48</td>
<td>99.83</td>
<td>99.99</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>n = 100</td>
<td>1.00</td>
<td>34.25</td>
<td>88.16</td>
<td>99.40</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>n = 20</td>
<td>5.00</td>
<td>18.69</td>
<td>33.27</td>
<td>48.71</td>
<td>62.80</td>
<td>73.75</td>
<td>82.46</td>
<td>88.24</td>
<td>91.70</td>
<td>95.19</td>
</tr>
<tr>
<td></td>
<td>n = 50</td>
<td>5.00</td>
<td>38.34</td>
<td>76.65</td>
<td>94.59</td>
<td>98.88</td>
<td>99.77</td>
<td>99.95</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>n = 100</td>
<td>5.00</td>
<td>78.08</td>
<td>99.43</td>
<td>99.98</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>n = 20</td>
<td>10.00</td>
<td>34.42</td>
<td>54.40</td>
<td>70.50</td>
<td>82.34</td>
<td>89.37</td>
<td>93.99</td>
<td>96.44</td>
<td>97.75</td>
<td>98.83</td>
</tr>
<tr>
<td></td>
<td>n = 50</td>
<td>10.00</td>
<td>61.55</td>
<td>92.01</td>
<td>98.79</td>
<td>99.80</td>
<td>99.96</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>n = 100</td>
<td>10.00</td>
<td>93.04</td>
<td>99.94</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
unit roots is entertained, with those obtained when the ADF tests are used to test for unit roots only at the zero frequency. The series are published only in semiannual form, and are:

These series are shown* in Fig. 6, and Table 9 presents the results of testing for a unit root only at the zero frequency, using the ADF tests. We allowed for drift and/or trend in the ADF regressions, and used the strategy of Dolado et al. (1990) to determine their final inclusion. The “augmentation level,” \( q \), for the regressions was chosen as the minimum required to obtain “clean” autocorrelation and partial autocorrelation functions for the associated residuals; and a seasonal dummy variable was included with any drift term. § As can be seen, the ADF tests suggest in each case that the series are \( I(1) \). Accordingly, one would infer that the series can be made stationary via simple first-differencing.

*In the case of series (4) the plotted series is for the number of cellphones, rather than the six-monthly increase in this number. The latter series is the one that is analyzed.

1In Table 9, \( t_0 \) denotes the ADF unit root “t-test” with drift and trend terms included in the fitted regression; \( F_{ut} \) is the corresponding ADF “F-test” for a unit root and zero trend; \( t_d \) is the unit root “t-test” with a drift but no trend in the fitted regression; \( F_{ud} \) is the corresponding “F-test” for a unit root and a zero drift; and \( t \) is the ADF unit root test when the fitted regression has no drift or trend term included. Finite-sample critical values for our “t-tests” and “F-tests” come from MacKinnon (1991) and from Dickey and Fuller (1979, 1981), respectively.

2The simulation results of Dods and Giles (1995) favor this approach, in terms of low pre-test size distortion, for sample sizes such as ours. In fact the results are not sensitive to the method used to choose this augmentation level.

3The need to include this dummy variable follows from the results of Ghysels et al. (1994). There was no evidence that any of the series other than series (4) may be \( I(2) \).
FIGURE 6  (a) New Zealand knitted fabric sales; (b) Canadian precast concrete price index (1981 = 100); (c) Canadian gas production; (d) Number of U.S. cellphones.
In Table 10 we show the results of applying the semiannual HEGY tests outlined in this chapter. In the cases 1, 3, and 4 we find that in fact the series are SI(1)—i.e., they each have unit roots at both the zero and π frequencies. So, in fact two-period differencing is the appropriate filter needed to make these series stationary. The earlier conclusion that series 2 has a unit root (only) at the zero frequency is upheld in Table 10, but these examples certainly illustrate the need for care when applying standard ADF tests to semiannual data.

6 CONCLUSIONS

In this chapter we have discussed a framework for testing for unit roots in time-series data that are reported at a maximum frequency of twice a year. In such cases, the possibility of unit roots at the zero and/or π frequencies arises, and the testing strategies developed by Hylleberg et al. (1990) and Beaulieu and Miron (1993) provide a natural framework to exploit. The asymptotic null distributions of the tests we propose are derived, discussed, and summarized, and in certain cases they can be linked to existing results associated with the well-known tests of Dickey and Fuller (1979, 1981). Simulated percentiles for the finite-sample null distributions of the test statistics are tabulated for various sample sizes and combinations of deterministic terms in the fitted regressions. The powers of the tests are simulated under a similar range of situations, and their application is illustrated with several actual semiannual time series.

The analysis in this chapter may be extended in several directions. The obvious one is to investigate the properties of corresponding tests for cointegration. Other topics of interest include the size robustness of the tests to autocorrelation in the data-generating process, to the method used to select an appropriate “augmentation level,” and to structural breaks in the data. These remain matters for future research.
<table>
<thead>
<tr>
<th>Fitted model&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Test</th>
<th>Series (1)</th>
<th></th>
<th>Series (2)</th>
<th></th>
<th>Series (3)</th>
<th></th>
<th>Series (4)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n(q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>66 (1)</td>
<td>32 (3)</td>
<td>22 (3)</td>
<td>22(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D, S, T</td>
<td></td>
<td>$t_{dt}$</td>
<td>-0.473</td>
<td>-0.439</td>
<td>-0.930</td>
<td>1.712</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{ut}$</td>
<td>1.824</td>
<td>2.281</td>
<td>2.009</td>
<td>10.194</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D, S, NT</td>
<td></td>
<td>$t_{d}$</td>
<td>-1.946</td>
<td>-1.900</td>
<td>l(1)</td>
<td>0.989</td>
<td>n.a.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$F_{ud}$</td>
<td>3.226</td>
<td>4.571</td>
<td>1.971</td>
<td>n.a.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND, NS, NT</td>
<td></td>
<td>t</td>
<td>1.837</td>
<td>l(1)</td>
<td>n.a.</td>
<td>1.848</td>
<td>l(1)</td>
<td>n.a.</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup><sup>n</sup> is the sample size; q is the “augmentation level” for the tests; D, S, T denotes “drift, seasonal dummy, trend”; D, S, NT denotes “drift, seasonal dummy, no trend”; ND, NS, NT denotes “no drift, no seasonal dummy, no trend.” The other notation is defined in the text.
**Table 10** Results of Seasonal Unit Root Tests

<table>
<thead>
<tr>
<th>Fitted model</th>
<th>Test</th>
<th>Series (1)</th>
<th>Series (2)</th>
<th>Series (3)</th>
<th>Series (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Statistic</td>
<td>Outcome</td>
<td>Statistic</td>
<td>Outcome</td>
</tr>
<tr>
<td>D, S, T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t₁</td>
<td></td>
<td>-0.439</td>
<td>Zero root</td>
<td>-0.473</td>
<td>Zero root</td>
</tr>
<tr>
<td>t₂</td>
<td></td>
<td>-2.389</td>
<td>π root</td>
<td>-5.072</td>
<td>π root</td>
</tr>
<tr>
<td>F₁₂</td>
<td></td>
<td>2.887</td>
<td>SI(1)</td>
<td>13.113</td>
<td></td>
</tr>
<tr>
<td>D, S, NT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t₁</td>
<td></td>
<td>-1.900</td>
<td>Zero root</td>
<td>-1.946</td>
<td>Zero root</td>
</tr>
<tr>
<td>t₂</td>
<td></td>
<td>-2.419</td>
<td>π root</td>
<td>-5.376</td>
<td>π root</td>
</tr>
<tr>
<td>F₁₂</td>
<td></td>
<td>4.896</td>
<td>SI(1)</td>
<td>20.866</td>
<td></td>
</tr>
<tr>
<td>D, NS, T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t₁</td>
<td></td>
<td>-0.405</td>
<td>Zero root</td>
<td>-0.464</td>
<td>Zero root</td>
</tr>
<tr>
<td>t₂</td>
<td></td>
<td>-1.103</td>
<td>π root</td>
<td>-5.152</td>
<td>π root</td>
</tr>
<tr>
<td>F₁₂</td>
<td></td>
<td>0.686</td>
<td>SI(1)</td>
<td>13.530</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Fitted model</th>
<th>Test</th>
<th>Statistic</th>
<th>Outcome</th>
<th>Statistic</th>
<th>Outcome</th>
<th>Statistic</th>
<th>Outcome</th>
<th>Statistic</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>D, NS, NT (q)</td>
<td>(t_1)</td>
<td>-1.896</td>
<td>Zero root</td>
<td>-2.003</td>
<td>Zero root</td>
<td>1.420</td>
<td>Zero root</td>
<td>0.837</td>
<td>Zero root</td>
</tr>
<tr>
<td></td>
<td>(t_2)</td>
<td>-1.145</td>
<td>(\pi) root</td>
<td>-5.458</td>
<td>(\pi) root</td>
<td>-0.510</td>
<td>(\pi) root</td>
<td>0.055</td>
<td>(\pi) root</td>
</tr>
<tr>
<td></td>
<td>(F_{12})</td>
<td>2.518</td>
<td>SI(1)</td>
<td>21.551</td>
<td>SI(1)</td>
<td>1.332</td>
<td>SI(1)</td>
<td>0.967</td>
<td>SI(1)</td>
</tr>
<tr>
<td>ND, NS, NT (q)</td>
<td>(t_1)</td>
<td>0.434</td>
<td>Zero root</td>
<td>1.837</td>
<td>Zero root</td>
<td>1.752</td>
<td>Zero root</td>
<td>1.276</td>
<td>Zero root</td>
</tr>
<tr>
<td></td>
<td>(t_2)</td>
<td>-1.181</td>
<td>(\pi) root</td>
<td>-6.440</td>
<td>(\pi) root</td>
<td>-0.748</td>
<td>(\pi) root</td>
<td>-0.130</td>
<td>(\pi) root</td>
</tr>
<tr>
<td></td>
<td>(F_{12})</td>
<td>0.799</td>
<td>SI(1)</td>
<td>38.202</td>
<td>SI(1)</td>
<td>1.935</td>
<td>SI(1)</td>
<td>1.718</td>
<td>SI(1)</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

The authors are grateful to Robert Draeseke, Judith Giles, Nilanjana Roy, Gugsa Werkneh, and participants in the University of Victoria Econometrics Colloquium, for their helpful comments on an earlier version of this work.

REFERENCES


Petroleum Monitoring Agency Canada, various years. Canadian Petroleum Industry Monitoring Reports.


Using Simulation Methods for Bayesian Econometric Models

John Geweke
University of Iowa, Iowa City, Iowa, U.S.A.

William McCausland
Université de Montréal, Montréal, Québec, Canada

John Stevens
Board of Governors of the Federal Reserve System, Washington, D.C., U.S.A.

1 INTRODUCTION

Subjective uncertainty is a central concept in economic theory and applied economics. In economic theory, it characterizes the beliefs of economic agents about the state of their environment. In applied economics, subjective uncertainty describes the situation of investigators who assess competing models based on their implications for what might be observed, and the circumstances of decision makers who must act, given limited information. With the application of the expected utility paradigm in increasingly richer environments, explicit distributional assumptions have become common, but closed-form analytical expressions for the distribution of observables are typically unobtainable. In this environment, simulation methods—the representation of probability distributions by related finite samples—have become important tools in economic theory.
In applied economics the possibility of proceeding strictly analytically is also remote. Even in the simplest typical situation the investigator or decision maker must proceed knowing the observables which are random variables in models of behavior, but not knowing the specification of tastes and technology that the theorist takes as fixed. Bayesian inference formalizes the applied economics problem in exactly this way: given a distribution over competing models and the prediction of each model for observables, the distribution of competing models conditional on the observables is well defined. However, the technical tasks in moving from even such well-specified models and data to the conditional distribution over models are more daunting than those found in economic theory. In the past decade very substantial progress has been made in the development of simulation methods suited to this task. Section 2 of this chapter reviews the conditional distributions of interest to the investigator or decision maker. Section 3 describes how modern simulation methods permit access to these distributions, using some simple examples to illustrate the methods. Section 4 sets forth recently developed numerical procedures for the explicit comparison of econometric models. The last section of the chapter takes up simple, effective ways of simultaneously realizing the rich promise of explicitly Bayesian methods and dealing with the desire of decision makers to change investigators’ assumptions at low cost.

Throughout, the chapter demonstrates the practical implementation of all these methods in the Bayesian analysis, computation, and communication (BACC) software. This embeds Bayesian tools in mathematical software such as Matlab and Gauss. It provides a seamless integration of commands for Bayesian inference with built-in general-purpose commands for computation, graphics, and flow control that are already familiar to many users. The BACC web site, http://www.cirano.gc.ca/~bacc, provides full details and the software itself.

2 BAYESIAN INFERENCE

This section provides a brief overview of Bayesian inference with reference to its application in economics. The purpose is to set the contribution of simulation methods in an explicit context of concepts and notation. Every attempt has been made to distill a large literature in statistics to what is essential to Bayesian inference as it is usually applied to economic problems. If this endeavor has been successful then this section also provides a sufficient introduction for econometricians with little or no grounding in Bayesian methods to appreciate some of the contributions, both realized and potential, of simulation methods to economic science.
Most of the material here is standard, reflecting much more comprehensive treatments including Jeffreys (1939, 1961), Zellner (1971), Berger (1985), Bernardo and Smith (1994), and Poirier (1995). At two junctures the exposition departs from the usual development. The first is the concept of a complete model (Section 2.1), which is the specification of a proper predictive distribution over an explicit class of events. This concept can be a clarifying analytical device.

The other deviation from the standard treatment is the decomposition of the marginal likelihood in terms of predictive densities (Section 2.3). This development was first provided explicitly by Geisel (1977), but has largely been ignored in the subsequent literature. The decomposition is the quantitative expression of the fact that predictive power is the scientifically relevant test of the validity of a hypothesis (Jeffreys, 1939; Friedman, 1953).

This review concentrates entirely on exact, finite sample methods. As is the case in non-Bayesian statistics, given suitable regularity conditions there exist useful asymptotic approximations to the exact, finite sample results. Bernardo and Smith (1994, Sect. 5.3) provide an accessible introduction to these results. Asymptotic methods are complementary to, rather than a prerequisite for, the posterior simulation methods taken up subsequently in Section 3.

2.1 Basic Concepts and Notation

Bayesian inference takes place in the context of one or more parametric econometric models. Let $y_t$ denote a $p \times 1$ vector of observable random vectors over a sequence of discrete time units $t = 1, 2, \ldots$. The history of the sequence $\{y_t\}$ at time $t$ is given by $Y_t = \{y_s\}_{s=1}^t \in \Psi$; $Y_0 = \{\emptyset\}$. A model, $A$, specifies a corresponding sequence of probability density functions $p(y_t | Y_{t-1}, \theta, A)$ in which $\theta$ is a $k \times 1$ vector of unknown parameters, $\theta \in \Theta \subset \mathbb{R}^k$, and $A$ denotes the model.* In this section we shall condition on a single model, but subsequently, in Section 2.3, several models will be entertained simultaneously.

The probability density function (p.d.f.) of $Y_T$, conditional on the model $A$ and parameter vector $\theta$, is $p(Y_T | \theta, A) = \prod_{t=1}^T p(y_t | Y_{t-1}, \theta, A)$. Conditional on observed $Y_T$, the likelihood function is any function

*Throughout, $p(\cdot)$ denotes a generic probability density function with respect to a measure $d\nu(\cdot)$ and $P(\cdot)$ a generic cumulative distribution function. The conditioning set makes clear the specific distribution or density intended. The measure $d\nu(\cdot)$ permits continuous, discrete, and mixed random variables.
L(θ; Y_T, A) ∝ p(Y_T|θ, A). If the model specifies that the \( y_t \) are independent and identically distributed then \( p(y_t|Y_{t-1}, \theta, A) = p(y_t|\theta, A) \) and \( p(Y_T|\theta, A) = \prod_{t=1}^{T} p(y_t|\theta, A) \). More generally, the index \( t \) may pertain to cross-sections, to time series, or both, but time-series models and language are used here for specificity.

If, in addition, the model \( A \) also provides the distribution of \( /C18 \), then it also provides the joint distribution of \( /C18 \) and \( Y_T \). In particular, if \( p(\theta|/C18) \) denotes the prior density, then

\[
p(Y_T, \theta|A) = p(\theta|A) \prod_{t=1}^{T} p(y_t|Y_{t-1}, \theta, A) = p(\theta|A)p(Y_T|\theta, A)
\]

(1)

but we also have

\[
p(Y_T, \theta|A) = p(\theta|Y_T, A)p(Y_T|A)
\]

(2)
in which

\[
p(Y_T|A) = \int_{\theta} p(Y_T|\theta, A)p(\theta|A) \, d\theta
\]

(3)
is the marginal likelihood\(^*\) of model \( A \). Therefore,

\[
p(\theta|Y_T, A) = p(Y_T|\theta, A)p(\theta|A)/p(Y_T|A) \propto p(Y_T|\theta, A)p(\theta|A)
\]
is the posterior density of \( \theta \) in model \( A \), so long as

\[
\int_{\theta} p(Y_T|\theta, A)p(\theta|A) \, d\theta = 1
\]
is absolutely convergent. This last condition is typically, but not necessarily, satisfied and easy to verify. For example, boundedness of the likelihood function \( p(Y_T|\theta, A) \) in \( \theta \) is sufficient, since \( \int_{\theta} p(\theta|A) \, d\theta = 1 \). However, if the likelihood function is unbounded, it is vital to confirm the absolute convergence of expression (4). Equations (1) and (2) are central, either explicitly or implicitly, to scientific learning. The former is used to express the reduction of reality to \( \theta \) inherent in the model \( A \), and the latter is used to learn about reality from the perspective of this particular simplification. This section outlines the basic principles of the explicit, or Bayesian, approach to learning.

In addition to the data density \( p(Y_T|\theta, A) \) and the prior density \( p(\theta|A) \), a model also specifies a density \( p(\omega|Y_T, \theta, A) \) for a vector of interest \( \omega \in \Omega \subseteq R^k \). This vector represents entities that the model is intended to describe. Whereas \( \theta \) is specific to \( A \), \( \omega \) remains the same across

\(^*\)This terminology dates at least to Raiffa and Schlaifer (1961, Sect. 2.1), which also treats these topics.
models. For example, suppose that a model specifies a first-order autoregressive process for \( y_t, \) \((y_t - \theta_1) = \theta_2(y_{t-1} - \theta_1) + \varepsilon_t \) with \( \varepsilon_t \sim N(0, \theta_3) \). If \( \omega' = (y_{T+1}, y_{T+2}) \), the first two postsample observations, then \( p(\omega'|Y_T, \theta, A) \) is a bivariate normal density with mean and variance:

\[
\begin{bmatrix}
\theta_1 + \theta_2(y_T - \theta_1) \\
\theta_1 + \theta_2^2(y_T - \theta_1)
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
1 & \theta_2 \\
\theta_2 & 1 + \theta_2^2
\end{bmatrix}
\]

respectively. If a second model specifies a second-order autoregressive process for \( y_t, \) \((y_t - \theta_1) = \theta_2(y_{t-1} - \theta_1) + \theta_3(y_{t-2} - \theta_1) + \varepsilon_t \) with \( \varepsilon_t \sim N(0, \theta_4) \), then \( p(\omega'|Y_T, \theta, A) \) is again a bivariate normal density, but with mean and variance:

\[
\begin{bmatrix}
\theta_1 + \theta_2(y_T - \theta_1) + \theta_3(y_{t-1} - \theta_1) \\
\theta_1 + (\theta_2^2 + \theta_3^2)(y_T - \theta_1)(y_T - \theta_1) + \theta_2\theta_3(y_{t-1} - \theta_1)
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
1 & \theta_2 \\
\theta_2 & 1 + \theta_2^2
\end{bmatrix}
\]

respectively. Since \( p(\omega'|Y_T, \theta, A) \) implies marginal distributions for subvectors of \( \omega \), one need not explicitly elaborate all of \( \omega \). Indeed, much scientific discourse can be interpreted as specification of \( \omega \). A complete model consists of three components: \( p(Y_T|\theta, A), p(\theta|A), \) and \( p(\omega|Y_T, \theta, A) \).

Without loss of generality, let the objective of inference when there is one model be

\[
E[h(\omega)|Y_T, A] \tag{5}
\]

for suitably chosen \( h(\cdot) \). This formulation includes several special cases of interest. If a hypothesis restricts \( \theta \) to a set \( \Theta_0 \), then, by taking \( h(\omega) = \chi_{\Theta_0}(\theta) \), we have \( E[h(\omega)|Y_T, A] = P(\theta \in \Theta_0|Y_T, A) \), the posterior probability that the hypothesis is true.\(^*\) To illustrate, consider the hypothesis that \( \{y_t\} \) is stationary in the example just presented. In the first-order autoregression, take \( \Theta = \{\theta_2; -1 < \theta_2 < 1\} \). In the second-order autoregression, take \( \Theta = \{\{\theta_2, \theta_3\); \(1 - \theta_2^2 - \theta_3^2 = 0 \Rightarrow \theta_3 > 1\} \). Here, and in general, nuisance parameters pose no particular difficulties.

Another important class of cases arises from prediction problems, \( \omega' = (y_{T+1}, \ldots, y_{T+p}) \). Through the appropriate choice of \( h(\omega) \) this category includes expected values, turning point probabilities, and

\(^*\)Here, and throughout, \( \chi_S(z) \) is the indicator function \( \chi_S(z) = 1 \) if \( z \in S \) and \( \chi_S(z) = 0 \) if \( z \notin S \).
predictive intervals. In the example just set forth, suppose that $y_{T-2} < y_{T-1} < y_T$. If a turning point at time $t$ is said to occur if $y_{t-2} < y_{t-1} < y_t > y_{t+1} > y_{t+2}$, then a turning point at time $T$ is the set of events $\Omega' = \{\omega: \omega_2 < \omega_1 < y_T\}$. Hence, for $h(\omega) = \chi_{\Omega'}(\omega)$, $E[h(\omega)|Y_T, A]$ is the probability of a turning point at time $T$, where $T$ is the end of the sample.

Yet another useful class of functions is $h(\omega) = L(a_1, \omega) - L(a_2, \omega)$, in which $L(a, \omega)$ denotes the loss incurred if action $a$ is taken and then the realization of the vector of interest is $\omega$. To examine a specific case, suppose that in our example $y_t$ is the logarithm of tax revenue in period $t$. A policy maker must either commit ($a_1$) or not commit ($a_2$) to a program that utilizes tax revenues in periods $T+1$ and $T+2$. Then, the policy maker’s loss function $L(a, \omega)$ might be monotone decreasing in $\omega_1 + \omega_2$ for $a = a_1$, and monotone increasing in $\omega_1 + \omega_2$ for $a = a_2$, and consequently $h(\omega)$ is monotone decreasing in $\omega_1 + \omega_2$. The solution of the decision problem is to commit to the project if $E[h(\omega)|Y_T, A] < 0$ and not commit if $E[h(\omega)|Y_T, A] > 0$.

The posterior moment, Eq. (5), can be expressed:

$$E[h(\omega)|Y_T, A] = \int_{\theta} \int_{\mathcal{R}} h(\omega)p(\theta|Y_T, A)p(\theta|Y_T, A) \, \text{d}v(\theta) \, \text{d}v(\theta)$$

$$= \int_{\theta} \int_{\mathcal{R}} h(\omega)p(\theta|Y_T, A)p^*(\theta|Y_T, A) \, \text{d}v(\theta) \, \text{d}v(\theta)$$

$$= \int_{\theta} p^*(\theta|Y_T, A) \, \text{d}v(\theta)$$

(6)

where $p^*(\theta|Y_T, A) \propto p(\theta|Y_T, A) \propto p(\theta|A)p(Y_T|\theta, A)$ is any posterior density kernel for $\theta$. It clearly matters not which posterior kernel is used. However, the problem of evaluating integrals—one in the numerator, the other in the denominator—remains paramount.

The importance of verifying the absolute convergence of the integral in the denominator of the right-hand side of Eq. (6) has already been noted. It is, of course, equally important to verify the absolute convergence of the numerator of Eq. (6). Together, both conditions are equivalent to the existence of the posterior moment Eq. (5). It is straightforward to verify these convergence conditions in the examples discussed above.

*More generally, any non-negative function proportional to a probability density is a density kernel.*
Many of these ideas can be illustrated in the standard linear model. For an observable $T \times 1$ vector of dependent variables $y$ and $T \times k$ matrix of fixed covariates $X$:

$$y = X\beta + \varepsilon; \quad \varepsilon | X \sim N(0, h^{-1}I_T); \quad \text{rank}(X) = k \tag{7}$$

The parameter $h$ is the precision of the i.i.d. disturbances, $\varepsilon_1, \ldots, \varepsilon_T$; it is the inverse of $\text{var}(\varepsilon_1) = \sigma^2$. Consider the independent prior distributions for $\beta$ and $h$:

$$\beta \sim N(\beta_0, H^{-1}) \tag{8}$$

$$\varepsilon^2 h \sim \chi^2(\nu) \tag{9}$$

where $H$ is a fixed precision matrix, $\beta$ is a fixed mean vector, and $\varepsilon^2$ and $\nu$ are fixed scalars. In any given application Eqs. (8) and (9) are not necessarily adequate expressions of prior beliefs. However, the specification in Eqs. (8) and (9) has attractive analytical properties that will become clear in due course. Moreover, in many cases it is straightforward to modify the posterior distribution implied by the prior distributions Eqs. (8) and (9), to express the posterior distributions corresponding to Eq. (7) and alternative prior distributions, using simple numerical methods described in Section 5.

From Eq. (8):

$$p(\beta) = (2\pi)^{-k/2} |H|^{1/2} \exp[-0.5(\beta - \beta_0)^T H(\beta - \beta_0)] \tag{10}$$

and from Eq. (9):

$$p(h) = [2^{\nu/2}\Gamma(\nu/2)]^{-1}(\varepsilon^2 h)^{\nu/2} h^{(\nu-2)/2} \exp(-\varepsilon^2 h/2) \tag{11}$$

Since Eq. (7) is equivalent to the conditional data density:

$$p(y|X, \beta, h) = (2\pi)^{-T/2} h^{T/2} \exp[-0.5h(y - X\beta)'(y - X\beta)] \tag{12}$$

*If instead $X$ is random with p.d.f. $p(X|\eta)$, $p(\beta, h, \eta|A) = p(\beta, h|A)p(\eta|A)$ and $p(\eta|x, \beta, h, \eta) = p(\eta|x, \beta, h)$, then $X$ is ancillary and the analysis that follows still pertains. For further discussion of ancillarity see Bernardo and Smith (1994, Sect. 5.1.4). The condition of weak exogeneity in the econometrics literature (Engle et al., 1983; Steel and Richard, 1991) is closely related.

*More generally, the precision of any random variable is the inverse of its variance. Nor is Eq. (7), necessarily. We return to this important question in greater depth in Sections 2.3 and 2.4.
a posterior density kernel is the product of Eqs. (10)–(12):

\[
(2\pi)^{-(T+k)/2}[2^{\nu/2}\Gamma(\nu/2)]^{-1} \times \left| H \right|^{1/2} (\nu/2) \times \left[ h^{T+\nu-2}/\nu \right] \exp(-s^2h/2) \times \exp\left[ -0.5[(\beta - \tilde{\beta})'H(\beta - \tilde{\beta}) + h(y - X\beta)'(y - X\beta)] \right]
\]

(13a)

\[
\times \left( \frac{H}{C_0} \right)^{1/2} \left( \frac{H}{C_0} + hX'X \right) = (\beta - \tilde{\beta})'H(\beta - \tilde{\beta}) + Q
\]

(13b)

\[
(\beta - \tilde{\beta})'H(\beta - \tilde{\beta}) + h(y - X\beta)'(y - X\beta) = (\beta - \tilde{\beta})'H(\beta - \tilde{\beta}) + Q
\]

(13c)

\[
\frac{1}{C_2H} = \frac{1}{C_2} \left( \frac{1}{C_1} \right)_a
\]

(13d)

To simplify this expression, complete the square in \(\beta\) of the term in brackets in Eq. (13d), yielding

\[
(\beta - \tilde{\beta})'H(\beta - \tilde{\beta}) + h(y - X\beta)'(y - X\beta) = (\beta - \tilde{\beta})'H(\beta - \tilde{\beta}) + Q
\]

where

\[
H = H + hX'X
\]

(14)

\[
\tilde{\beta} = H^{-1}(H\beta + hX'y) = H^{-1}(H\beta + hX'b)
\]

(15)

\[
Q = hy'y + \beta' \tilde{H}\beta - \tilde{\beta}'H\tilde{\beta}
\]

\[
= h\nu s^2 + (b - \tilde{\beta})'hX'X(b - \tilde{\beta}) + (\beta - \tilde{\beta})'H(\beta - \tilde{\beta})
\]

(16)

with \(b\) denoting the coefficients in the ordinary least squares fit of \(y\) to \(X\), \(b = (XX)^{-1}X'y\); \(s^2 = (y - Xb)'(y - Xb)/v\) and \(v = T - k\). If Eq. (13) is interpreted as a function of \(\beta\) only, that function must be a posterior density kernel for \(\beta\) conditional on \(h\), and our square completion shows that \(p(\beta|h,y,X) \propto \exp\{-0.5(\beta - \tilde{\beta})'H(\beta - \tilde{\beta})\}. Consequently,

\[
\beta|h,y,X \sim N(\tilde{\beta}, \tilde{H}^{-1})
\]

(17)

Interpreting Eq. (13) as a function of \(h\) alone:

\[
p(h|\beta,y,X) \propto h^{T+\nu-2}/\nu \exp\{-0.5[s^2 + (y - X\beta)'(y - X\beta)]h]\}
\]

and consequently,

\[
[s^2 + (y - X\beta)'(y - X\beta)]h|\beta,y,X \sim \chi^2(T + \nu)
\]

(18)

The distributions in Eqs. (8) and (9) are special cases of conditionally conjugate priors (to be defined shortly). They are attractive because they lead to the tractable results [Eqs. (17) and (18)]. Yet these results are not directly useful, for they do not provide distributions conditional only on the data and prior information. However, they form the basis of an attractive simulation method discussed in Section 3.3.
2.2 Conjugate Prior Distributions

The prior distribution \( p(\theta | A) \) is a representation of belief in the context of model \( A \). In selecting a prior or data distribution, the richer the class of functional forms from which to choose the more adequate the representation of prior beliefs possible. On the other hand, the choice is constrained by the tractability of the posterior density \( p(\theta | Y_T, A) \propto p(\theta | A)p(Y_T | \theta, A) \), which is jointly determined by the choice of functional forms for the data density and prior density. The search for rich tractable classes of prior distributions may be formalized by considering classes of prior densities, \( p(\theta | A) = p(\theta | \gamma, A) \), where \( \gamma \) is a parameter vector that indexes prior beliefs. We have used this approach in the linear model: e.g., the prior distribution and exponential family of distributions (which includes the standard linear model) the search for rich tractable classes of prior distributions may be formalized by considering classes of prior densities, \( p(\theta | A) = p(\theta | \gamma, A) \), where \( \gamma \) is a parameter vector that indexes prior beliefs.

Indeed, one can begin with this property as the definition of conjugate; see Berger (1985, Sect. 4.2.2) and Poirier (1995, Sect. 6.7). The definition here is that used by Bernardo and Smith (1994, Sect. 5.2.1) and Zellner (1971, Sect. 2.3). For the exponential family of distributions (which includes the standard linear model) the two are equivalent [see Bernardo and Smith (1994), Proposition 5.4].
\[ f_{\theta_1}(s_{\gamma_j}^*) = \gamma_j^*(j = 2, \ldots, r^*)|\theta_1, \theta_2 = \theta_2^0, A) d\theta_1 < \infty \]  
and \[ p(\theta_1|s_{\gamma_j}, A) \propto p(\theta^*_j) = \gamma_j^*(j = 2, \ldots, r^*)|\theta_1, \theta_2 = \theta_2^0, A). \]

The prior distributions Eqs. (8) and (9), are conditionally conjugate, but not conjugate, in the linear model, Eq. (7). In this example, the prior density for \( \theta' = (\beta', h) \) is indexed by \( \gamma = (\beta, \mathbf{H}, \mathbf{s}^2, \mathbf{y}) \). In the linear model, because

\[ p(y|X, \beta, h) \propto h^{T/2} \exp\{-0.5h[\mathbf{s}^T \mathbf{s} + (\beta - \mathbf{b})' \mathbf{X}' \mathbf{X} (\beta - \mathbf{b})]\} \tag{19} \]

the vector \( \mathbf{s}_T = [T, \mathbf{h}, \mathbf{s}^2, \mathbf{X}' \mathbf{X}] \) is a sufficient statistic.* Conditioning on \( h = h_0, \)
\[ p(y|X, \beta) \propto \exp\{-0.5(\beta - \mathbf{b})' \mathbf{H}(\beta - \mathbf{b})\}. \]
Since \( p(\beta) \propto \exp\{-0.5(\beta - \beta_0)' \mathbf{H}(\beta - \beta_0)\}, \) the prior density, Eq. (10), is conditionally conjugate.

Likewise, conditioning on \( \beta = \beta_0, \) \( p(y|X, h) \propto h^{T/2} \exp(-\mathbf{s}^2 h/2) \) where \( \mathbf{s}^2 = \mathbf{s}^T \mathbf{s} + (\beta_0 - \mathbf{b})' \mathbf{X}' \mathbf{X}(\beta_0 - \mathbf{b}) \). Hence, the prior density, Eq. (11), \( p(h) \propto h^{(T-2)/2} \exp(-\mathbf{s}^2 h/2) \), is conditionally conjugate.

### 2.3 Model Comparison and Combination

Often one has under consideration several complete models, say \( A_1, \ldots, A_J \):

\[ p(\theta_j|A_j) (\theta_j \in \Theta_j), \quad p(Y_T|\theta_j, A_j), \quad p(\omega|Y_T, \theta_j, A_j) \quad (j = 1, \ldots, J) \]

The numbers of parameters in the models need not be the same, and various models may or may not nest one another. If we assign prior probabilities \( P(A_j) (j = 1, \ldots, J) \) to the respective models, with \( \sum_{j=1}^{J} P(A_j) = 1 \), then there is a complete probability structure for \( \{A_j, \theta_j\}_{j=1}^{J}, Y_T, \) and \( \omega \). There is no essential conceptual distinction between model and prior, since one could just as well regard the entire collection as the model, with \( \{P(A_j), p(\theta_j|A_j)\}_{j=1}^{J} \) as the characterization of the prior distribution. At an operational level the distinction is usually clear and useful in that one may undertake the essential computations one model at a time.

Suppose that the posterior moment, Eq. (5), is ultimately of interest. The formal solution is

\[ E[h(\omega)|Y_T] = \sum_{J=1}^{J} E[h(\omega)|Y_T, A_j]P(A_j|Y_T) \tag{20} \]

---

*This follows from the Neyman factorization criterion [Bernardo and Smith, 1994, Sect.4.5.2]. Less formally, from Eq. (19) it is clear that we only need to know \( \mathbf{s}_T \) to write the likelihood function for \( \beta \) and \( h \).
which is known as model averaging. Clearly, $E[h(\omega)|Y_T, A_j]$ is given by Eq. (6) with $A = A_j$. There is nothing new in this part of Eq. (20). From Bayes’ rule:

$$P(A_j|Y_T) = P(A_j)p(Y_T|A_j)/\sum_{j=1}^{J} P(A_j)p(Y_T|A_j)$$

$$= P(A_j)\int_{\theta_j} p(Y_T|\theta_j, A_j)p(\theta_j|A_j)\,d\theta_j/\sum_{j=1}^{J} P(A_j)p(Y_T|A_j)$$

$$\propto P(A_j)\int_{\theta_j} p(Y_T|\theta_j, A_j)p(\theta_j|A_j)\,d\theta_j = P(A_j)p(Y_T|A_j)$$

(21)

where $p(Y_T|A_j) = \int_{\theta_j} p(Y_T|\theta_j, A_j)p(\theta_j|A_j)\,d\theta_j$ is the marginal likelihood of model $j$, consistent with the definition in Eq. (3). Notice it is important that the properly normalized prior and properly normalized data density, and not arbitrary kernels of these densities, be used in forming the marginal likelihood.

Model averaging thus involves three steps. First, obtain the posterior moments, Eq. (6), corresponding to each model. Second, obtain the relative values of $P(A_j|Y_T)$ from Eq. (21). Finally, obtain the posterior moment using Eq. (20), which now only involves simple arithmetic, recognizing that $\sum_{j=1}^{J} P(A_j|Y_T) = 1$. Variation of the prior model probabilities $P(A_j)$ is a trivial step, as is the revision of the posterior moment following the introduction of a new model or deletion of an old one from the conditioning set of models.

From Eq. (21), for any pair of models $A_j$ and $A_k$:

$$P(A_j|Y_T)/P(A_k|Y_T) = [P(A_j)/P(A_k)][p(Y_T|A_j)/p(Y_T|A_k)]$$

(22)

This ratio of probabilities is the posterior odds ratio in favor of model $j$ versus model $k$. It is invariant with respect to the addition and deletion of models from the set $\{A_j\}_{j=1}^{J}$ under consideration, so long as the prior probabilities $\{P(A_j)\}_{j=1}^{J}$ are changed in a logically consistent fashion—i.e., ratios $P(A_j)/P(A_k)$ remain unchanged for all included models.* The posterior odds ratio is expressed in Eq. (22) as the product of the prior odds ratio in favor of model $j$ versus model $k$, $P(A_j)/P(A_k)$, and the Bayes factor in favor of model $j$ versus model $k$, $p(Y_T|A_j)/p(Y_T|A_k)$.

---

*This property is analogous to the independence of irrelevant alternatives in the qualitative choice literature; see Poirier (1997).
In the case of the standard linear model it is straightforward to work out the marginal likelihood and Bayes factors if $h$ is fixed. The product of the properly normalized prior and data densities is

$$p(\beta)p(y|X, \beta, h) = (2\pi)^{-\left(T+k\right)/2} h^{T/2} |H|^{1/2} \exp\left\{-0.5 \times [h(y - X\beta)'(y - X\beta) + (\beta - \tilde{\beta})' H(\beta - \tilde{\beta})]\right\}$$

(23)

The term in brackets may be expressed as

$$\left(\beta - \tilde{\beta}\right)' H(\beta - \tilde{\beta}) + Q$$

(24)

with $\tilde{\beta}$, $H$ and $Q$ as defined in Eqs. (14)–(16). Substituting Eq. (24) into (23), the marginal likelihood is

$$\int_{\beta} p(\beta)p(y|X, \beta, h) \, d\beta$$

$$= (2\pi)^{-\left(T+k\right)/2} h^{T/2} |H|^{1/2} \int_{\beta} \exp\left\{-0.5 \times [\beta - \tilde{\beta}]' H(\beta - \tilde{\beta}) + Q\right\} \, d\beta$$

$$= (2\pi)^{-T/2} h^{T/2} |H|^{1/2} \exp\left(-Q/2\right)$$

$$= \sum_{n=1}^{N} \exp\left(-0.5[hv]\right)$$

(25)

From the last expression it is apparent that the marginal likelihood of a linear model depends on more than the least squares fit of $y$ to $X$, which is measured by the sum of squared residuals $\sum_i y_i^2$. It also depends on the squared Euclidean distance of the least squares fit $b$ from the posterior mean $\tilde{\beta}$, using the data-based norm $hX'X$; the squared distance of the prior mean $\beta$ from the posterior mean $\tilde{\beta}$, using the prior-based norm $H$; and the fraction of posterior precision accounted for by prior precision, as measured by $|H|/|H|^\ast$.

### 2.4 Hierarchical Priors and Latent Variables

A *hierarchical prior distribution* expresses the prior in two or more steps. The two-step case specifies a model:

$$p^{(1)}(Y_T|\theta, \lambda, A)$$

(26)

*Equation (21) shows that the marginal likelihood of model $j$, $p(Y_T|A_j)$, is the measure of how well model $A_j$ predicted the observed data $Y_T$ that is relevant for the comparison of model $j$ with any other models. For a formal development of this idea, see Geweke (2001).
with a prior density for $\theta \in \Theta$ conditional on a vector of hyperparameters $\phi \in \Phi$:  
\[ p^{(2)}(\theta|\phi, A) \]  
and a prior density for $\phi$ and $\lambda \in \Lambda$:  
\[ p^{(3)}(\phi, \lambda|A) \]  
it being understood in Eq. (26) that $p^{(1)}(Y_T|\theta, \lambda, A) = p(Y_T|\theta, \lambda, \phi, A)$.  
The full prior density for all parameters and hyperparameters is  
\[ p(\theta, \phi, \lambda|A) = p^{(3)}(\phi, \lambda|A)p^{(2)}(\theta|\phi, A) \]  
There is no fundamental difference between this prior density and the one described in Section 2.1, since  
\[ p(\theta, \lambda) = \int_\phi p^{(2)}(\theta|\phi, A)p^{(3)}(\phi, \lambda|A) d\phi \]  
However, the hierarchical formulation is often so convenient as to render fairly simple the analysis of posterior densities that would otherwise be quite difficult. Given a hierarchical prior, one may express the full posterior density:  
\[ p(\theta, \phi, \lambda|Y_T, A) \propto p^{(1)}(Y_T|\theta, \lambda, A)p^{(2)}(\theta|\phi, A)p^{(3)}(\phi, \lambda|A) \]  
A latent variable model expresses the likelihood function in two or more steps. In the two-step case the likelihood function may be written:  
\[ p^{(1)}(Y_T|Z_T^*, \lambda, A) \]  
where $Z_T^* \in Z_T$ is a matrix of latent variables and $\lambda \in \Lambda$. The model for $Z_T^*$ is  
\[ p^{(2)}(Z_T^*|\phi, A) \]  
and the prior density for $\phi \in \Phi$ and $\lambda$ is  
\[ p^{(3)}(\phi, \lambda|A) \]  
The full prior density for all parameters and unobservable variables is  
\[ p(Z_T^*, \phi|Y_T, A) = p^{(3)}(\phi, \lambda|A)p^{(2)}(Z_T^*|\phi, A) \]  
and the full posterior density is  
\[ p(Z_T^*, \phi|Y_T, A) \propto p^{(1)}(Y_T|Z_T^*, \lambda, A)p^{(2)}(Z_T^*|\phi, A)p^{(3)}(\phi, \lambda|A) \]  
Comparing Eqs. (26)–(30) with (31)–(35), it is apparent that the latent variable model is formally identical to a model with a two-stage hierarchical prior, the latent variables corresponding to the intermediate
level of the hierarchy. With appropriate marginalization of Eq. (35) one may obtain \( p(Z_T^T|Y, A) \), which fully reflects uncertainty about the parameters. If one is interested only in \( \lambda \) and \( \phi \), these distributions may also be obtained by marginalization of Eq. (35). Marginalization requires integration over \( Z_T^T \), which is possible analytically only in special cases. If the problem is approached using the simulation methods described beginning in the Section 3, then this integration simply amounts to discarding simulated values of \( Z_T^T \).

A simple example of a latent variable model is provided by the textbook probit model:

\[
y^* = X\beta + \epsilon, \quad \epsilon|X \sim N(0, I), \quad \text{rank}(X) = k, \quad d_i = x_{(t, \infty)}(y_i^*) \quad (36)
\]

in which the \( T \times k \) matrix of covariates \( X = [x_1', \ldots, x_T']' \) and decision vector \( d' = (d_1, \ldots, d_T) \) are observed, but \( y^{**} = (y_1^*, \ldots, y_T^*) \) is latent. To complete the model take

\[
\beta \sim N(\beta, \mathbf{H}^{-1}) \quad (37)
\]

In the equivalent formulation of this model using a hierarchical prior the parameter vector is \((y^*, \beta)\). The first level of the hierarchical prior is \( \beta \sim N(\beta, \mathbf{H}^{-1}) \), corresponding to \( p^{(3)} \) with \( \phi = \beta \). The second level is \( y^*|\beta, X) \sim N(\beta, I) \), with \( \theta = y^* \) in the hierarchical prior interpretation and \( Z_T^T = y^* \) in the latent variable interpretation. (There is no analog of \( \lambda \) in this example.) The data distribution is

\[
p(d|y^*) = \prod_{t=1}^T \left[ x_{(0, \infty)}(y_t^*) d_t + \chi_{(-\infty, 0)}(y_t^*) (1 - d_t) \right] \nonumber
\]

Either formulation leads to the same joint distribution for \( \beta, y^*, \) and \( d \):

\[
p(\beta, y^*, d|X) = (2\pi)^{-\left( T+k \right)/2} |\mathbf{H}|^{1/2} \exp\left[ -0.5 (\beta - \beta)^T \mathbf{H} (\beta - \beta) \right] \\
\times \prod_{t=1}^T \left[ x_{(0, \infty)}(y_t^*) d_t + \chi_{(-\infty, 0)}(y_t^*) (1 - d_t) \right] \quad (38)
\]

The main conceptual point is that since Bayesian inference conditions on the observables \( (d, X) \), parameters and latent variables have the same standing as unknown entities whose joint distribution with the observables is given by the model. As we shall see in in Section 3.3, this formulation provides a basis for computations as well.
3 POSTERIOR SIMULATION METHODS

The objective of inference in a single model:

$$E[h(\omega)|Y_T, A] = \int_{\Theta} \int_{\Omega} h(\omega)p(\omega|\theta, Y_T, A)p(\theta|Y_T, A) \, d\nu(\omega) \, d\nu(\theta),$$

can be evaluated analytically only in a few specific simple cases. This section describes simulation methods for obtaining a sequence of strongly consistent approximations to $E[h(\omega)|Y_T, A]$, and Section 4 will take up the process of model averaging. In most applications, it is generally straightforward to find a function $g(Y_T, \theta)$, possibly random, with the property:

$$E[g(Y_T, \theta)|Y_T, A] = E[h(\omega)|Y_T, A] = \int_{\Theta} h(\omega)p(\omega|Y_T, \theta, A) \, d\omega = \tilde{g}(y_T, \theta) \quad (39)$$

Finding this function is trivial if $h(\omega)|Y_T, A$ is deterministic.* If $h(\omega)$ is random, then it is often straightforward to take $\omega \sim p(\omega|Y_T, \theta, A)$ and then $g(Y_T, \theta) = h(\omega)$. In either case,

$$E[h(\omega)|Y_T, A] = \int_{\Theta} \tilde{g}(Y_T, \theta)p(\theta|Y_T, A) \, d\nu(\theta) = \tilde{g}$$

More generally, one may be able to find a function satisfying Eq. (39), but for which

$$\text{var}[g(Y_T, \theta)|Y_T, A] < \text{var}[h(\omega)|Y_T, A] \quad (40)$$

The turning point example of Section 2.1 provides an illustration. Recall that in this example the objective was to evaluate $P(y_{T+2} < y_{T+1} < y_T|Y_T)$ and to this end we took $\omega' = (y_{T+1}, y_{T+2})$. One could draw $\omega'(Y_T, \theta)$ and use the random function $g(Y_T, \theta) = h(\omega) = \chi_{\Omega^*}(\omega)$. Alternatively, one could draw only $\omega_1'(Y_T, \theta)$ and use the random function $g(Y_T, \theta) = P(\omega_2 < \omega_1|Y_T, \theta)$, which requires only the ability to evaluate the univariate standard normal c.d.f. Yet a third alternative is to employ the deterministic function $g(Y_T, \theta) = P(\omega_2 < \omega_1 < y_T|Y_T, \theta)$ using bivariate quad-

*The evaluation of $g(Y_T, \theta)$ may not be trivial at all. For example, Bajari (1997) has functions of interest whose evaluation requires the solution of a system of nonlinear differential equations.
In each case \( \text{E}[g(Y_T, \theta)Y_T, \theta] = P(y_{T+2} < y_{T+1} < y_T|Y_T, \theta) \) but \( \text{var}[g(Y_T, \theta)Y_T, \theta] \) is greatest in the first alternative, less in the second, and zero in the third.*

Throughout we shall make use of the notation \( g(Y_T, \theta) \), it always being implicit that Eq. (39) is satisfied.

If one could also make a sequence of independent draws \( \{-\{\theta^m\}\}_{m=1}^\infty \) from the posterior distribution, then by choosing \( \omega^m \sim p(\omega|Y_T, \theta^m, A) \) one could guarantee \( M^{-1} \sum_{m=1}^M h(\omega^m) \overset{a.s.}{\rightarrow} \text{E}[h(\omega)|Y_T, A] \). However, direct simulation from the posterior distribution is rarely possible. This section describes methods for obtaining a sequence \( \{-\{\theta^m\}\}_{m=1}^\infty \), and an associated weighting function \( w(\omega) \), with the property that if \( \text{E}[h(\omega^m)|\theta^m] = \text{E}[h(\omega)|Y_T, \theta^m, A] \) for a corresponding sequence \( \{-\{\theta^m\}\}_{m=1}^\infty \), then

\[
\sum_{m=1}^M w(\theta^m)h(\omega^m) / \sum_{m=1}^M w(\theta^m) \quad \overset{a.s.}{\rightarrow} \quad \text{E}[h(\omega)|Y_T, A]
\]

The ability to generate such sequences has improved greatly in the past 10 years, due in large part to the development of Markov chain Monte Carlo (MCMC) methods and the dramatic decrease in the cost of computing. We begin by reviewing two more established methods, acceptance and importance sampling, and then move on to the Gibbs sampler and the Hastings–Metropolis algorithm as examples of MCMC. This is followed by a more abstract development of MCMC theory, a description of some of the hybrid procedures that make MCMC a powerful tool for posterior simulation, and a discussion of the evaluation of approximation error. The section concludes with a description of some public domain software for posterior simulation and two simple examples. The emphasis here is on concepts and practicality; only references to proofs of theorems are given. A more general and extensive introduction is provided by Gelman et al. (1995). A concise presentation of the relevant continuous state space Markov chain theory that underlies MCMC procedures is given in Tierney (1994).

### 3.1 Acceptance Sampling

Acceptance sampling is the algorithm that underlies the generation of random variables from most familiar univariate distributions like the normal and the gamma (Press et al., 1992). The idea behind acceptance

*In some cases the left-hand side of Eq. (40) can be made quite small indeed, and asymptotically it may be made to approach zero (Geweke, 1988).
sampling is to generate a random vector* from a distribution that is similar, in an appropriate sense, to the posterior distribution, and then to accept that drawing with a probability that depends on the drawn value of the vector. If this acceptance probability function is chosen correctly, then the accepted values will have the desired distribution.

Theorem 1
Suppose that \( p^*(\theta|Y_T, A) \) is any kernel of the posterior density \( p(\theta|Y_T, A) \). Let \( s^*(\theta) \) be a source density kernel with respect to the same measure \( d\nu(\theta) \) as \( p(\theta|A) \), with support \( S \) and the property:

\[
0 \leq p^*(\theta|Y_T, A)/s^*(\theta) \leq a < \infty \quad \forall \theta \in \Theta
\]  

(41)

Suppose that at each iteration a candidate \( \bar{\theta} \) is drawn from the source density, and \( \theta^{(m)} \) is set to \( \bar{\theta} \) with probability \( p^*(\bar{\theta}|Y_T, A)/[as^*(\bar{\theta})] \). This process is repeated until \( \theta^{(m)} \) is set to a candidate \( \theta \). Then \( \theta^{(m)} \overset{iid}{\sim} p(\theta|Y_T, A) \).

Proof
See Geweke (1999, Theorem 3.1.1).

A successful application of acceptance sampling has three requirements. First, there must be a source density corresponding to a distribution from which it is efficient and convenient to make i.i.d. draws. Second, there must be a known upper bound on the ratio of the posterior density to the source density. Finally, the frequency with which \( \theta^{(m)} \) set to \( \bar{\theta} \) must not be so small that the whole algorithm is impractical. The upper bound must be established analytically, whereas efficiency can be evaluated through experimentation. Notice that draws from the source density may (and usually do) involve acceptance sampling: e.g., if the source density is a normal or gamma density, the software used to draw from this density very likely employs acceptance sampling, a fact typically transparent to the software user.

Acceptance sampling produces an i.i.d. sequence \( \{\theta^{(m)}\} \). It follows from the strong law of large numbers that \( \bar{g}_M = M^{-1} \sum_{m=1}^{M} g(Y_T, \theta^{(m)}) \overset{a.s.}{\longrightarrow} \bar{g} \). If in addition \( \sigma^2 = \text{var}[g(Y_T, \theta)|Y_T, A] \) exists then,

*We ignore the distinction between the mathematical properties of a sequence of random variables and the properties of (what is properly called) a pseudo-random variable sequence created using a computer. For a discussion of these issues see Geweke (1996) and references therein.
from the Lindberg–Levy central limit theorem, 

\[ M^{1/2} [\tilde{g}_m - \bar{g}] \overset{d}{\longrightarrow} N(0, \sigma^2), \]

and a second application of the strong law of large numbers yields

\[ \tilde{\sigma}^2 = M^{-1} \sum_{m=1}^{M} [g(Y_T, \theta^{(m)}) - \bar{g}_m] \overset{a.s.}{\longrightarrow} \sigma^2. \]

Thus, if the posterior variance of the function of interest exists, a central limit theorem may be used in the usual way to assess the numerical accuracy of the approximation of 

\[ E[h(\omega)|Y_T, A] \]

by 

\[ M^{-1} \sum_{m=1}^{M} g(Y_T, \theta^{(m)}). \]

### 3.2 Importance Sampling

Rather than accept only a fraction of the draws from the source density, it is possible to retain all of them, and consistently approximate the posterior moment by appropriately weighting the draws. The probability density function of the source distribution is then called the importance sampling density, a term due to Hammersly and Handscomb (1964), who were among the first to propose the method. It appears to have been introduced to the econometrics literature by Kloek and van Dijk (1978).

To help distinguish between acceptance and importance sampling we shall indicate the importance sampling distribution by its density \( j(\theta) \) with respect to the same measure \( d \theta \) as the prior density \( p(\theta|Y_T, A) \). Let \( k(\theta) \) be any kernel of \( j(\theta) \), and let \( p^*(\theta|Y_T, A) \) be any kernel of \( p(\theta|Y_T, A) \).

**Theorem 2**

Suppose \( E[g(Y_T, \theta)|Y_T, A] \) exists, and the support of \( j(\theta) \) includes \( \Theta \). Then

\[ \tilde{g}_M = \frac{\sum_{m=1}^{M} g(Y_T, \theta^{(m)}) w(\theta^{(m)})}{\sum_{m=1}^{M} w(\theta^{(m)})} \overset{a.s.}{\longrightarrow} E[g(Y_T, \theta)|Y_T, A] = \bar{g} \]

where \( w(\theta) = p^*(\theta|Y_T, A) \) is the corresponding weighting function. If in addition both \( E[w(\theta)|Y_T, A] = \int_\Theta w(\theta) p(\theta|Y_T, A) d\theta \) and \( \text{var}[g(Y_T, \theta)|Y_T, A] \) exist, then

\[ M^{1/2} (\tilde{g}_M - \bar{g}) \overset{d}{\longrightarrow} N(0, \sigma^2) \tag{42} \]

and

\[ \tilde{\sigma}_M^2 = M \sum_{m=1}^{M} [g(Y_T, \theta^{(m)}) - \bar{g}_M]^2 w(\theta^{(m)})^2 / \left[ \sum_{m=1}^{M} w(\theta^{(m)}) \right]^2 \overset{a.s.}{\longrightarrow} \sigma^2 \tag{43} \]
Proof

See Geweke (1989b, Theorems 1 and 2).

This result provides a practical way to assess approximation error and also indicates conditions in which the method of importance sampling will work well. Small variance in \( w(\theta) \), perhaps reflecting close upper and lower bounds on \( w(\theta) \), will lead to small values of \( \sigma^2 \) relative to \( \text{var}[g(Y_T, \theta)|Y_T, A] \). Of course, the existence of \( \mathbb{E}[w(\theta)|Y_T, A] \) and \( \text{var}[g(Y_T, \theta)|Y_T, A] \) must be verified analytically. The following implication of Corollary 2 is often useful in the latter undertaking.

Corollary 1

If \( \text{var}[g(Y_T, \theta)|Y_T, A] \) exists and the weighting function \( w(\theta) = p^*(\theta|Y_T, A)/j^*(\theta) \) is bounded, then Eqs. (42) and (43) are true.

The hypothetical special case \( j(\theta) \propto p(\theta|Y_T, A) \) corresponds to i.i.d. sampling from the posterior distribution, since the weighting function is then constant. In this case, \( \sigma^2 = \text{var}[g(Y_T, \theta)|Y_T, A] \), which can serve as a benchmark in evaluating the adequacy of \( j(\theta) \) in all other cases. The ratio \( \text{var}[g(\theta, Y_T)|Y_T, A]/\sigma^2 \) has been termed the relative numerical efficiency of the importance sampling approximation to \( \mathbb{E}[g(Y_T, \theta)|Y_T, A] \) (Geweke, 1989b): it indicates the ratio of iterations using \( p(\theta|Y_T, A) \) itself as the importance sampling density, to the number using \( j(\theta) \), required to achieve the same accuracy of approximation of \( \bar{g} \). Since both the numerator and denominator of the ratio \( \text{var}[g(\theta, Y_T)|Y_T, A]/\sigma^2 \) can be approximated consistently as the number of draws \( M \) increases, this is a practical indication of the computational efficiency of importance sampling. Relative numerical efficiency much less than 1.0 (less than 0.1, certainly less than 0.01) indicates poor imitation of \( p(\theta|Y_T, A) \) by \( j(\theta) \), possibly the existence of a better importance sampling distribution or the failure of the underlying convergence conditions for Eq. (43).

Importance sampling is an important useful tool in modifying prior distributions. Suppose that models \( A_1 \) and \( A_2 \) are distinguished only by their prior densities \( p(\theta|A_j), j = 1, 2 \). Suppose that one has available an i.i.d. sample from the posterior density \( p(\theta|Y_T, A_1) \propto p(\theta|A_1)p(Y_T|\theta, A_1) \). If \( p(\theta|A_2)/p(\theta|A_1) \) is bounded above, then \( p(\theta|Y_T, A_1) \) is an importance sampling density for \( p(\theta|Y_T, A_2) \) that satisfies the conditions of Corollary 1. The weighting function is \( w(\theta) = p(\theta|A_2)/p(\theta|A_1) \). Thus, one may change the prior distribution without reworking the entire problem. The ability to do so makes conditionally conjugate prior distributions, of the kind discussed in Section 2 in conjunction with the standard linear model, attractive as reporting devices because an investigator’s results, produced
with such priors, may be modified by a client with different priors. This idea will be developed more fully in Section 5.

3.3 The Gibbs Sampler

The Gibbs sampler is an algorithm that has been used with noted success in many econometric models. It is one example of a wider class of procedures known as Markov chain Monte Carlo (MCMC). In these procedures the idea is to construct a Markov chain with state space \( \Theta \), and unique invariant distribution \( p(\theta | Y_T, A) \). Following an initial transient or burn-in phase, simulated values from the chain are used to approximate \( E[g(Y_T, \theta)|Y_T, A] \). The Gibbs sampler in the form described here was first fully developed by Gelfand and Smith (1990).

The Gibbs sampler begins with a partition, or blocking, of \( \theta, \theta' = (\theta_1', \ldots, \theta_B') \). In applications, the blocking is chosen so that it is possible to draw from each of the conditional p.d.f.'s, \( p(\theta_{(a)}|Y_T, \theta_{(a)} (a < b), \theta_{(b)} (a > b), A) \). This blocking can arise naturally, if the prior distributions for the \( \theta_{(b)} \) are independent and each is conditionally conjugate. To motivate the key idea underlying the Gibbs sampler, suppose—contrary to fact—that there existed a single drawing \( \theta^{(0)}, \theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_B^{(0)}) \), from \( p(\theta | Y_T, A) \). Successively make drawings from the conditional distributions as follows:

\[
\begin{align*}
\theta_1^{(1)} & \sim p(Y_T | \theta_1^{(0)}, \ldots, \theta_B^{(0)}, A) \\
\theta_2^{(1)} & \sim p(Y_T | \theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_B^{(0)}, A) \\
& \vdots \\
\theta_B^{(1)} & \sim p(Y_T | \theta_1^{(0)}, \ldots, \theta_B^{(0)}, A) \\
& \vdots \\
\theta_B^{(1)} & \sim p(Y_T | \theta_1^{(0)}, \ldots, \theta_B^{(0)}, A).
\end{align*}
\]

This defines a transition process from \( \theta^{(0)} \) to \( \theta^{(1)} = (\theta_1^{(1)}, \ldots, \theta_B^{(1)}) \). Since \( \theta^{(0)} \sim p(\theta | Y_T, A) \):

\[
(\theta_1^{(1)}, \ldots, \theta_B^{(1)}, \theta_1^{(0)}, \ldots, \theta_B^{(0)}) \sim p(\theta | Y_T, A)
\]

at each step in Eq. (44) by definition of the conditional density. In particular, \( \theta_1^{(1)} \sim p(\theta | Y_T, A) \).
Iteration of this algorithm produces a sequence \( \theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(m)}, \ldots, \) which is a realization of a Markov chain with probability density function

\[
K_G(\theta^{(m)}, \theta^{(m+1)}) = \prod_{b=1}^B \Pr\left[ \theta^{(m+1)}(b) \mid Y_T, \theta^{(m)}(a > b), \theta^{(m+1)}(a < b), A \right]
\]

Any single iterate \( \theta^{(m)} \) retains the property that it is drawn from the posterior distribution. For the Gibbs sampler to be practical, it is essential that the blocking be chosen in such a way that one can make the drawings in an efficient manner. In econometrics the blocking is often natural and the conditional distributions familiar. In making the drawings, Eq. (44), acceptance sampling is often useful.

The appeal of the Gibbs sampler is easy to illustrate with the standard linear model, Eqs. (7)–(9): The results, Eqs. (17) and (18), indicate that the blocking \( \theta_{(1)} = \beta, \theta_{(2)} = h \) meets the criterion that drawings can be made in an efficient manner. The probit model introduced in Section 2.4 is a further example, as noted in Albert and Chib (1993). From Eqs. (36) and (37) it is evident that conditional on the vector of latent variables \( y^* \) the distribution of \( \beta \) is given by Eq. (17) if we use \( y^* \) in place of \( y \) and set \( h = 1 \). Examination of the kernel of Eq. (38) in \( y^* \) shows that, given \( \beta \) and \( X \), the \( y^*_t \) are conditionally independent, with \( y^*_t \sim N(\beta' x_t, 1) \) truncated to \([0, \infty)\) if \( d_t = 1 \) and truncated to \((\infty, 0)\) if \( d_t = 0 \). An efficient algorithm for drawing from truncated normal distributions is given in Geweke (1991). In both cases, given drawings for the parameters it is straightforward to produce numerical approximations to \( \mathbb{E}[h(\omega) | Y_T, A] \), as indicated at the start of this section, and, as discussed in Section 2.1, the evaluation of \( \mathbb{E}[h(\omega) | Y_T, A] \) subsumes most of the uses to which these models are put.

Of course, if it really were possible to make an initial draw from the posterior distribution, then independence Monte Carlo would also be possible. An important remaining task is to elucidate conditions for the distribution of \( \theta^{(m)} \) to converge to the posterior for any \( \theta^{(0)} \in \Theta \). This is not trivial, because even if \( \theta^{(0)} \) were drawn from \( p(\theta | Y_T, A) \), the argument just given establishes only that any single \( \theta^{(m)} \) is also drawn from the posterior distribution. It does not establish that a single sequence \( \{\theta^{(m)}\}_{m=1}^\infty \) is representative of the posterior distribution. For example, if \( \Theta \) consists of two disjoint subsets \( \Theta_1 \) and \( \Theta_2 \) with \( \theta_1 > \theta_2 \), \( \forall \theta_1, \theta_2 \in \Theta_j \), then a Gibbs sampler that begins in \( \Theta_1 \) and has individual elements of \( \Theta \) as blocks will never visit \( \Theta_2 \) and vice versa. This situation clearly does not arise in the Gibbs samplers for the standard linear and probit models just described, but evidently a careful development of conditions under which \( \{\theta^{(m)}\} \)
converges in distribution to the posterior distribution is needed. We outline these developments in Section 3.5.

3.4 The Hastings–Metropolis Algorithm

The Hastings–Metropolis algorithm begins with an arbitrary transition probability density function \( q(x, y) \) indexed by \( x \in \Theta \) and with density argument \( y \in \Theta \), and with an arbitrary starting value \( \theta(0) \in \Theta \). The random vector \( \theta^* \) generated from \( q(\theta^m, \theta^*) \) is a candidate value for \( \theta^{m+1} \). The algorithm actually sets \( \theta^{m+1} = \theta^* \) with probability:

\[
\alpha(\theta^m, \theta^*) = \min \left\{ \frac{p(\theta^* | Y_T, A)q(\theta^*, \theta^m)}{p(\theta^m | Y_T, A)q(\theta^m, \theta^*)}, 1 \right\}
\]

\[
= \min \left\{ \frac{p(\theta^* | Y_T, A)/q(\theta^m, \theta^*)}{p(\theta^m | Y_T, A)/q(\theta^m, \theta^m)}, 1 \right\}
\]

(46)

otherwise, the algorithm sets \( \theta^{m+1} = \theta^m \). This defines a Markov chain with a generally mixed continuous–discrete transition probability from \( \theta^m \) to \( \theta^{m+1} \) given by

\[
K_H(\theta^m, \theta^{m+1}) = \begin{cases} 
q(\theta^m, \theta^{m+1})\alpha(\theta^m, \theta^{m+1}) & \text{if } \theta^{m+1} \neq \theta^m \\
1 - \int_\Theta q(\theta^m, \theta)\alpha(\theta^m, \theta) \, d\nu(\theta) & \text{if } \theta^{m+1} = \theta^m 
\end{cases}
\]

This form of the algorithm is due to Hastings (1970). The Metropolis et al. (1953) form takes \( q(\theta^m, \theta^*) = q(\theta^*, \theta^m) \). A simple variant that is often useful is the independence chain (Tierney, 1994), whereby \( q(\theta^m, \theta^*) = k(\theta^*) \). Then

\[
\alpha(\theta^m, \theta^*) = \min \left\{ \frac{p(\theta^* | Y_T, A)k(\theta^m)}{p(\theta^m | Y_T, A)k(\theta^*)}, 1 \right\} = \min \left\{ \frac{w(\theta^*)}{w(\theta^m)}, 1 \right\}
\]

where \( w(\theta) = p(\theta | Y_T, A)/k(\theta) \). The independence chain is closely related to acceptance sampling and importance sampling. In acceptance sampling, if the posterior density is low (high) relative to the source density the probability of acceptance is low (high). In importance sampling, if the posterior density is low (high) relative to the importance sampling the weight assigned to the draw is low (high). In the independence chain, to the extent the posterior density is lower (higher) relative to the proposal than was the case in the previously accepted draw, the probability of accepting the proposed vector is lower (one).

There is a simple two-step argument that motivates the convergence of the sequence \( \{\theta^m\} \) generated by the Hastings–Metropolis algorithm to
the posterior. (This approach is due to Chib and Greenberg, 1995.) First, observe that if the transition probability function \( p(\theta^{(m)}, \theta^{(m+1)}) \) satisfies the reversibility condition:

\[
p(\theta^{(m)}) p(\theta^{(m)}, \theta^{(m+1)}) = p(\theta^{(m+1)}) p(\theta^{(m)}, \theta^{(m)})
\]

(47)

for stated \( p(\cdot) \), then it has \( p(\cdot) \) as an invariant distribution. To see this, note that if Eq. (47) holds then

\[
\int_{\theta} p(\theta^{(m)}) p(\theta^{(m)}, \theta^{(m+1)}) d\nu(\theta^{(m)}) = \int_{\theta} p(\theta^{(m+1)}) p(\theta^{(m)}, \theta^{(m)}) d\nu(\theta^{(m)})
\]

\[
= p(\theta^{(m+1)}) \int_{\theta} p(\theta^{(m)}, \theta^{(m)}) d\nu(\theta^{(m)})
\]

\[
= p(\theta^{(m+1)})
\]

For \( \theta^{(m+1)} = \theta^{(m)} \), Eq. (47) is satisfied trivially. For \( \theta^{(m+1)} \neq \theta^{(m)} \), suppose without loss of generality that \( p(\theta^{(m+1)})/q(\theta^{(m)}, \theta^{(m+1)}) > p(\theta^{(m)})/q(\theta^{(m+1)}, \theta^{(m)}) \). Then

\[
p(\theta^{(m)}, \theta^{(m+1)}) = q(\theta^{(m)}, \theta^{(m+1)})
\]

and

\[
p(\theta^{(m+1)}, \theta^{(m)}) = q(\theta^{(m+1)}, \theta^{(m)}) \cdot \frac{p(\theta^{(m)})/q(\theta^{(m+1)}, \theta^{(m)})}{p(\theta^{(m+1)})/q(\theta^{(m)}, \theta^{(m+1)})}
\]

\[
= p(\theta^{(m)}) q(\theta^{(m)}, \theta^{(m+1)}) / p(\theta^{(m+1)})
\]

whence Eq. (47) is satisfied.

In implementing the Hastings–Metropolis algorithm the transition probability density function must share two important properties. First, it must be possible to generate \( \theta^* \) efficiently from \( q(\theta^{(m)}, \theta^*) \). A second key characteristic of a satisfactory transition process is that the unconditional acceptance rate not be so low that the time required to generate a sufficient number of distinct \( \theta^{(m)} \) is too great.

In the case of the independence chain the Hastings–Metropolis algorithm will be efficient under essentially the same conditions that the corresponding importance sampling algorithm with the same \( j(\theta) \) will be efficient. If there are values of \( \theta \) for which \( p^*(\theta|Y_T, A)/j(\theta) \) is very much greater than at other values, then the importance sampling algorithm will place very high weights on these values, which are drawn infrequently relative to \( p^*(\theta|Y_T, A) \). The Hastings–Metropolis independence chain will tend to remain at such values for many successive iterations. In either case relative numerical efficiency will, as a consequence, be low.

The Hastings–Metropolis algorithm and the Gibbs sampling algorithm can be combined in a way that has proved quite useful in
states, so that distribution of accepted candidates. Assume that $K$ has no absorbing
forward to sample from $p(\theta_{(2)}|Y_T, \theta_{(1)}, A)$, but the distribution corresponding
to $p(\theta_{(2)}|Y_T, \theta_{(1)}, A)$ is intractable. The Hastings–Metropolis algorithm
can be used in these circumstances, and it often provides an efficient
solution to the problem. In what has become known as the Metropolis-
within-Gibbs procedure, at the $(m+1)$th iteration first draw $\theta_{(2)}^{(m+1)}$ from a
proposal density $q(\theta_{(2)}|\theta_{(1)}^{(m+1)})$. Accept this draw with probability
\[
\min \left\{ \frac{p(\theta_{(1)}^{(m+1)}, \theta_{(2)}^{(m+1)}|Y_T, A)}{p(\theta_{(1)}^{(m+1)}, \theta_{(2)}^{(1)}|Y_T, A)}, 1 \right\}
\]
If $\theta_{(2)}^{(m+1)}$ is accepted then $\theta_{(2)}^{(m+1)} = \theta_{(2)}^{(m)}$, and if not then $\theta_{(2)}^{(m+1)} = \theta_{(2)}^{(1)}$. The
extension of this procedure to multi-block Gibbs samplers, with a
Hastings–Metropolis algorithm used at some (or even all) of the blocks is
clear. This procedure is generally known as the Metropolis-within-Gibbs
algorithm. For further discussion see Zeger and Karim (1991) and Chib
and Greenberg (1995), and for a proof that the posterior distribution is an
invariant state of this Markov chain see Chib and Greenberg (1996).

3.5 Some MCMC Theory

Much of the treatment here draws heavily on the work of Tierney (1994),
who first used the theory of general state space Markov chains to
demonstrate convergence, and Roberts and Smith (1994), who elucidated
sufficient conditions for convergence that turn out to be applicable in a
wide variety of problems in econometrics.

Let $\{\theta^{(m)} \}_{m=0}^{\infty}$ be a Markov chain defined on $\Theta \subseteq \mathbb{R}^k$ with transition
density $K: \Theta \times \Theta \rightarrow \mathbb{R}^+$ such that, for all $\nu$-measurable $\Theta_0 \subseteq \Theta$,
\[
P(\theta^{(m)} \in \Theta_0 | \theta^{(m-1)}) = \int_{\Theta_0} K(\theta^{(m-1)}, \theta) \, d\nu(\theta) + \nu(\theta^{(m-1)} \setminus \Theta_0) \chi_{\Theta_0}(\theta^{(m-1)}),
\]
where $\nu(\theta^{(m-1)} \setminus \Theta_0) = 1 - \int_{\Theta_0} K(\theta^{(m-1)}, \theta) \, d\nu(\theta)$.

The transition density $K$ is substochastic: it defines only the
distribution of accepted candidates. Assume that $K$ has no absorbing
states, so that $r(\theta) < 1 \ \forall \ \theta \in \Theta$. The corresponding substochastic kernel
over $m$ steps is then defined iteratively,
\[
K^{(m)}(\theta^{(0)}, \theta^{(m)}) = \int_{\Theta} K^{(m-1)}(\theta^{(0)}, \theta) K(\theta, \theta^{(m)}) \, d\nu(\theta)
+ K^{(m-1)}(\theta^{(0)}, \theta^{(m)}) \nu(\theta^{(m)}) + [r(\theta^{(0)})]^{m-1} K(\theta^{(0)}, \theta^{(m)}).
\]
This describes all $m$-step transitions that involve at least one accepted move. As a function of $\theta^{(m)}$ it is the p.d.f. with respect to $\nu$ of $\theta^{(m)}$, excluding realizations with $\theta^{(m)} = \theta^{(0)} \forall n = 1, \ldots, m$. For any $\nu$-measurable $\Theta_0$ let $P^{(m)}(\theta^{(0)}, \Theta_0)$ denote the $m$th iterate of $P$:

$$P^{(m)}(\theta^{(0)}, \Theta_0) = \int_{\Theta_0} K^{(0)}(\theta^{(0)}, \theta) \, d\nu(\theta) + (r(\theta^{(0)}))^m \chi_{\Theta_0}(\theta^{(0)})$$

An invariant distribution of the transition density $K$ is a function $p(\theta)$ that satisfies

$$P(\Theta_0) = \int_{\Theta_0} p(\theta) \, d\nu(\theta) = \int_{\Theta_0} P(\theta^{(m)} \in \Theta_0 | \theta^{(m-1)} = \theta) p(\theta) \, d\nu(\theta)$$

for all $\nu$-measurable $\Theta_0$. Let $\Theta^* = \{ \theta \in \Theta : p(\theta) > 0 \}$. The density $K$ is $p$-irreducible if for all $\theta^{(0)} \in \Theta^*$, $P(\Theta_0) > 0$ implies that $P^{(m)}(\theta^{(0)}, \Theta_0) > 0$ for some $m \geq 1$.

The transition density $K$ is aperiodic if there exists no $\nu$-measurable partition $\Theta = \bigcup_{r \geq 1} \Theta_r$ such that

$$P(\theta^{(m)} \in \tilde{\Theta}_{m \text{mod}(r)} | \theta^{(0)} = \omega) = 1 \forall m$$

It is Harris recurrent if $P(\theta^{(m)} \in \Theta_0 \text{ i.o.} | \theta^{(0)}) = 1$ for all $\nu$-measurable $\Theta_0$ with $\int_{\Theta_0} p(\theta) \, d\nu(\theta) > 0$ and all $\theta^{(0)} \in \Theta$. It follows directly that if a kernel is Harris recurrent, then it is $p$-irreducible. A kernel whose invariant distribution is proper, and that is both aperiodic and Harris recurrent, is ergodic by definition (Tierney, 1994, pp. 1712–1713).

A useful metric in what follows is the total variation norm for signed and bounded measures $\mu$ defined over the field of all $\nu$-measurable sets $S_\omega$ on $\Theta : ||\mu|| = \sup_{\theta^{(0)} \in S_\omega} \mu(\Theta_0) - \inf_{\theta^{(0)} \in S_\omega} \mu(\Theta_0)$.

**Theorem 3**

Convergence of Continuous State Markov Chains

Suppose $p(\theta | Y_T, A)$ is an invariant distribution of the transition density $K(\theta, \theta^*)$.

(A) If $K$ is $p(\theta | Y_T, A)$-irreducible, then $p(\theta | Y_T, A)$ is the unique invariant distribution.

---

*The expression “i.o.” in $P(\theta^{(m)} \in \tilde{\Theta}_0 \text{ i.o.} | \theta^{(0)}) = 1$ means “infinitely often.” The condition is that $\lim_{M \to \infty} P(\sum_{m=1}^{M} \chi_{\Theta_0}(\theta^{(m)}) \leq L) = 0 \forall L$.}
If \( K \) is \( p(\theta | Y_T, A) \)-irreducible and aperiodic, then except possibly for \( \theta^{(0)} \) in a set of posterior probability 0, \( \| P^{(0)}(\theta^{(0)}, \cdot) - P(\cdot | Y_T, A) \| \to 0 \). If \( K \) is ergodic (i.e., it is also Harris recurrent) then this occurs for all \( \theta^{(0)} \).

If \( K \) is ergodic with invariant distribution \( p(\theta | Y_T, A) \), then for all \( g(Y_T, \theta) \) absolutely integrable with respect to \( p(\theta | Y_T, A) \) and for all \( \theta^{(0)} \in \Theta \),

\[
M^{-1} \sum_{m=1}^{M} g(Y_T, \theta^{(m)}) \xrightarrow{a.s.} \int_{\Theta} g(Y_T, \theta)p(\theta | Y_T, A) d\nu(\theta).
\]

Proof

(A) and (B) follow immediately from Theorem 1 and (C) from Theorem 3 in Tierney (1994).

For the Gibbs sampling algorithm we argued informally in Section 3.3 that \( p(\theta | Y_T, A) \) is an invariant distribution. More formally, from Eq. (45) we have for the blocking \( \theta^{(0)} = (\theta^{(1)}, \theta^{(2)}) \):

\[
K_{G}(\theta, \theta')p(\theta | Y_T, A) d\nu(\theta)
\]

\[
= \int_{\Theta} p(\theta^{(1)} | Y_T, \theta^{(2)}, A)p(\theta^{(2)} | Y_T, \theta^{(1)}, A)p(\theta | Y_T, A) d\nu(\theta)
\]

\[
= p(\theta^{(1)} | Y_T, \theta^{(2)}, A) \int_{\Theta} p(\theta^{(2)} | Y_T, \theta^{(1)}, A)p(\theta | Y_T, A) d\nu(\theta)
\]

\[
= p(\theta^{(1)} | Y_T, \theta^{(2)}, A)p(\theta^{(2)} | Y_T, \theta^{(1)}, A) = p(\theta^{(1)} | Y_T, \theta^{(2)})
\]

The general result for more than two blocks follows by induction. Thus, it is the uniqueness of the invariant state that is at issue in establishing convergence of the Gibbs sampler. The following result is immediate and is often easy to apply.

Corollary 2

A Sufficient Condition for Convergence of the Gibbs Sampler

Suppose that for every point \( \theta^{*} \in \Theta \) and every \( \Theta_0 \subseteq \Theta \) with the property \( P(\theta \in \Theta_0 | Y_T, A) > 0 \), it is the case that \( P_{G}(\theta^{(m+1)} \in \Theta_0 | Y_T, \theta^{(m)} = \theta^{*}, A) > 0 \), where \( P_{G}(\cdot) \) is the probability measure induced by the Gibbs sampler. Then the Gibbs transition kernel is ergodic.

Proof

The conditions ensure that \( P_{G} \) is aperiodic and absolutely continuous with respect to \( p(\theta | Y_T, A) \). The result follows from Corollary 1 of Tierney (1994).
Tierney (1994) and Roberts and Smith (1994) show that the convergence properties of the Hastings–Metropolis algorithm are inherited from those of $q(\theta, \theta^*)$: if the Markov chain with transition probability $q$ is aperiodic and $p(\theta|Y, A)$-irreducible, then so is the Markov chain simulated by the Hastings–Metropolis algorithm. This feature leads to a sufficient condition for convergence analogous to Corollary 2.

**Theorem 4**

A Sufficient Condition for Convergence of the Hastings–Metropolis Algorithm

Suppose that for every point $\theta^* \in \Theta$ and every $\Theta_0 \subseteq \Theta$ with the property $P(\theta \in \Theta_0|Y, A) > 0$, it is the case that $\int_{\Theta_0} q(\theta, \theta^*) \alpha(\theta, \theta^*) \mathrm{d}v(\theta^*) + r(\theta) \chi_{\Theta_0}(\theta) > 0$. Then, the Hastings–Metropolis density $K(\theta, \theta^*) = q(\theta, \theta^*) \alpha(\theta, \theta^*)$ is ergodic.

**Proof**

The conditions ensure that the transition kernel is aperiodic and that $p(\theta^*|Y, A)$ irreducible. Thus, by Corollary 2 of Tierney (1994), the Hastings–Metropolis density is Harris recurrent. Since the kernel is both aperiodic and Harris recurrent, it is ergodic.

The conditions in Corollary 2 and Theorem 4 are sufficient but not necessary. However, they are easier to apply directly than is Theorem 3. For some alternative sufficient conditions, see Geweke (1999).

### 3.6 Assessing Numerical Accuracy in Markov Chain Monte Carlo

In any practical application one is concerned with the discrepancy $\tilde{g}_M - \tilde{g}$. A leading analytical tool for assessing this discrepancy is a central limit theorem, if one can be obtained. This was accomplished in Section 3.1 for i.i.d. sampling from the posterior distribution, and in Section 3.2 for importance sampling. The assumption of independence, key to those results, does not apply in MCMC. The weaker assumption of uniform ergodicity yields a central limit theorem, however. Let $P^{(m)}(\theta^{(0)}, \Theta_0)$ denote $P(\theta^{(m)} \in \Theta_0|\theta^{(0)})$ for any $\theta^{(0)} \in \Theta$ and for any $\Theta_0 \subseteq \Theta$ for which $P(\theta \in \Theta_0|Y, A)$ is defined. The Markov chain is uniformly ergodic if

$$\sup_{\theta \in \Theta} \|P^{(m)}(\theta, \cdot) - P(\cdot|Y, A)\| \leq M^m$$

for some $M > 0$ and some positive $r < 1$.

Tierney (1994, p. 1714) demonstrates two results that are quite useful in establishing uniform ergodicity. First, an independence Metropolis kernel $j(\theta)$ with bounded weighting function $w(\theta) = p(\theta|Y)|j(\theta)$ is
uniformly ergodic. (Recalling the similarity between the independence Metropolis kernel and importance sampling and Corollary 1 this result is not surprising.) Second, if one kernel in a mixture of kernels is uniformly ergodic, then the mixture kernel itself is uniformly ergodic.

The interest in uniform ergodicity stems from the following central limit theorem. Note how close this result is to Corollary 1.

Theorem 5
A Central Limit Theorem for Markov Chain Monte Carlo

Suppose \( \{\theta^{(m)}\} \) is uniformly ergodic with equilibrium distribution \( p(\theta|Y_T, A) \). Suppose further that \( E[g(Y_T, \theta)|Y_T, A] = \bar{g} \) and \( \text{var}[g(Y_T, \theta)|Y_T, A] \) exist and are finite, and let \( \bar{g}_M = M^{-1} \sum_{m=1}^{M} g(Y_T, \theta^{(m)}) \). Then there exists finite \( \sigma^2 \) such that

\[
M^{1/2}(\bar{g}_M - \bar{g}) \xrightarrow{d} N(0, \sigma^2)
\]

Proof
Tierney (1994, Theorem 5), attributed to Cogburn [1972, Corollary 4.2(ii)].

Thus, for any Markov chain \( \{\theta^{(m)}\} \) with invariant distribution \( p(\theta|Y_T, A) \), one can guarantee Eq. (48) by mixing the chain with an independence Metropolis kernel and a bounded weighting function, so long as the posterior mean and variance are known to exist. If the likelihood function is bounded, then the prior distribution itself will provide such an independence transition kernel.

Nevertheless, some practical concerns remain. One difficulty is that useful conditions sufficient for approximation of the unknown constant \( \sigma^2 \) have not yet been developed. That is, there is no \( \hat{\sigma}^2_M \) for which \( \hat{\sigma}^2_M \rightarrow \sigma^2 \)
as there is for independence and importance sampling. A second difficulty is assessing the sensitivity of \( \theta^{(m)} \) to the initial condition \( \theta^{(0)} \). For example, in the case of a multimodal distribution, transitions between the neighborhoods of all the modes could be quite slow.

There is an extensive literature on this problem. A good introduction is provided by the papers of Gelman and Rubin (1992) and Geyer (1992) and their discussants. Geweke (1992) developed a consistent estimator of \( \sigma^2 \) in Eq. (48), under the strong condition that conventional time series mixing conditions (e.g., Hannan, 1970, pp. 207–210) apply to \( \{\theta^{(m)}\} \). There is no analytical foundation for this assumption, but these methods are now widely used and have proven reliable in the sense that they predict well the behavior of the Markov chain when it is restarted with a new initial condition, in econometric models.
In practice, some robustness to initial conditions is achieved by discarding initial iterations; 10 to 20% is common. By drawing $\theta^{(0)}$ from the prior distribution, using a random number generator with a fresh seed each time, several runs may provide some indication of whether the results are sensitive to initial conditions as they might be, e.g., given near-reducibility of the kind that may arise from severe multimodality. A formal test for sensitivity to initial conditions was developed by Gelman and Rubin (1992). For other tests for sensitivity to initial conditions see Geweke (1992) and Zellner and Min (1995).

### 3.7 Software

The BACC software provides its users with tools for Bayesian analysis, computation, and communication. Dynamically linked libraries (DLLs) implement these tools as extensions to popular mathematical applications such as Matlab, Splus, and Gauss. The DLLs and full source code are available for free from the website www.cirano.gc.ca/~bacc.* The BACC user has available, within the mathematical application, commands for various Bayesian tools. The `minst` command allows a user to set up a model instance, by identifying one of the available models, and specifying the known quantities of the model, such as data and fixed hyperparameters. The `postsim` command generates a posterior simulation structure for a model instance, from which the user can extract the simulated parameter vectors $\theta^{(m)}$, the evaluations $\log(p(\theta^{(m)}|A))$ of the log prior, and the evaluations $\log(p(Y_T|\theta^{(m)}, A))$ of the log likelihood. The `expect1` and `expectN` commands calculate estimates of moments and their standard errors. These commands are described in detail below. The `mlike` command computes estimates of the marginal likelihood for a model instance. This command is described in detail in Section 4.5.

The BACC user also has available all of the mathematical application’s built-in commands for matrix computation, graphics, program flow control, and I/O. The utility of this flexibility is illustrated in Section 5.2, where standard Matlab commands are used to calculate estimates of posterior moments not envisioned by the investigator, and to

---

*Complete documentation for all software is provided at the website. Since this software will continue to be developed and improved, some details provided in this chapter will become outdated. Users should rely on the website documentation for actual use rather than the descriptions in this chapter, which are intended to provide concrete examples of how Bayesian inference, development, and communication can proceed.
compute estimates of posterior moments for a client whose prior differs from that of the investigator.

The `expect1` command takes as input any (possibly weighted) sample, and calculates estimates of the mean and standard deviation. Users can also request estimates of the numerical standard error of the mean, for as many values of a taper parameter $L$ (discussed below) as they care to provide.

We describe the computations performed by the `expect1` command for the special case where the sample is the sequence \( \{g(Y_T, \theta^{(m)})\} \) of evaluations of a function of interest at the draws \( \theta^{(m)} \) of the parameter vector, and the weighting sequence is \( \{w^{(m)}\} \). The draws \( \theta^{(m)} \) and weights \( w^{(m)} \) represent the posterior distribution of the parameter vector.

Suppose there are Monte Carlo interactions, of which the first \( r \) are discarded as burn-in iterations. Then numerical approximation of the posterior mean of the function of interest is

\[
\tilde{g} = \frac{\sum_{m=r+1}^{M} w(\theta^{(m)})g(Y_T, \theta^{(m)})}{\sum_{m=r+1}^{M} w(\theta^{(m)})}
\]

The numerical approximation of the posterior standard deviation of the function of interest is

\[
\sqrt{\frac{\sum_{m=r+1}^{M} w(\theta^{(m)})[g(Y_T, \theta^{(m)}) - \tilde{g}]^2}{\sum_{m=r+1}^{M} w(\theta^{(m)})}}^{1/2}
\]

Different variants of a numerical standard error for the accuracy of the approximation of the true posterior moment \( E[g(Y_T, \theta)|Y_T, A] \) by the numerical approximation \( \tilde{g} \) are provided in BACC. To describe these, let \( n_M = (M-r)^{-1} \sum_{m=r+1}^{M} w(\theta^{(m)})g(\theta^{(m)}, Y_T) \) denote the numerator of \( \tilde{g} \) and let \( d_M = (M-r)^{-1} \sum_{m=r+1}^{M} w(\theta^{(m)}) \) denote the denominator of \( \tilde{g} \). Using the conventional asymptotic expansion (“delta method”):

\[
\text{var}(n_M/d_M) \approx \begin{bmatrix} d_M^{-1} - & n_Md_M^-2 \\ var(n_M) & \text{cov}(n_M, d_M) \\ \text{cov}(d_M, n_M) & var(d_M) \end{bmatrix} \begin{bmatrix} d_M^{-1} \\ -n_Md_M^-2 \end{bmatrix}
\]

The variants of the numerical standard error, \( s.e.(n_M/d_M) = [\text{var}(n_M/d_M)]^{1/2} \), are based on different approximations of \( \text{var}(n_M), \text{cov}(n_M, d_M) \) and \( \text{var}(d_M) \).
The first method assumes no serial correlation in \( \{\theta^m\} \), and is appropriate for independence or importance sampling. Following Geweke (1989b), and consistent with Eq. (43), this leads to

\[
\text{var}(n_M/d_M) \approx \sum_{m=r+1}^{M} w(\theta^m)^2 [g(\theta^m, Y_T) - \bar{g}]^2 / \left[ \sum_{m=r+1}^{M} w(\theta^m) \right]^2
\]

(The square root of the value on the right-hand side is reported by expect1.)

In the remaining methods, \( \text{var}(n_M) \), \( \text{cov}(n_M, d_M) \) and \( \text{var}(d_M) \) are approximated using conventional time-series methods for a wide sense stationary process similar to those described in Geweke (1992). In the case of \( \text{var}(d_M) \):

\[
\text{var}(d_M) \approx (M - r)^{-1} \sum_{s=-L+1}^{L-1} [(L - |s|)/L]c(s) \tag{49}
\]

where for \( s \geq 0, c(s) = c(-s) = (M - r)^{-1} \sum_{m=r+s+1}^{M} [w(\theta^m) - d_M]w(\theta^{(m-s)}) - d_M] \). For \( \text{var}(n_M) \), the approximation is the right hand side of Eq. (49) but with

\[
c(s) = c(-s) = (M - r)^{-1} \sum_{m=r+s+1}^{M} [g(\theta^m)w(\theta^m) - n_M]
\]

\[
\times [g(\theta^{(m-s)})w(\theta^{(m-s)}) - n_M]
\]

where \( s \geq 0 \). For \( \text{cov}(n_M, d_M) \) the approximation is the right-hand side of Eq. (49) but with

\[
c(s) = (M - r)^{-1} \sum_{m=\max(r+1, s+1)}^{\min(M, M-s)} [g(\theta^m)w(\theta^m) - n_M][w(\theta^{(m-s)}) - d_M]
\]

The remaining variants differ in the value of \( L \) chosen. One value of the numerical standard error is reported for each value of \( L \) chosen.*

The `expectN` command takes as input several weighted samples. It provides estimates of the numerical standard error for the combined sample, and conventional test statistics for the equality of the moments of the component samples, under the assumption that the samples are independent from one another.

*Small values of \( L \) assume a more rapid rate of decay in the autocovariance function of \( [g(Y_T, \theta^m)] \). In practice results are usually about the same for the three values. Substantial differences indicate that serial correlation may persist across a substantial fraction of the iterations, and a longer simulation may be warranted.
It computes the same variants of the numerical standard error as `expect1`. If there are $J$ samples, the combined posterior moment approximation is $\tilde{g} = \sum_{j=1}^{J} v_j \tilde{g}_j / \sum_{j=1}^{J} v_j$, where $\tilde{g}_j$ is the mean of the $j$th sample and $v_j$ is the inverse square of its numerical standard error. The `expectN` command provides the conventional chi-square test of $\tilde{g}_1 = \cdots = \tilde{g}_J$ (in all variants), and the marginal significance level of this test statistic. If the $J$ samples are the same posterior functions of interest, created using $J$ independent initial conditions for the same posterior simulator, then this test is essentially the convergence test proposed by Gelman and Rubin (1992).

Evidence of different values of the moments from different posterior samples is an indication that there may be sensitivity to starting values, or—almost equivalently—that an insufficient number of burn-in iterations were taken in approximating the moments.

### 3.8 Examples

Two examples illustrate the use of these methods, and will be used in subsequent portions of this chapter as well.* The first example is based on the hedonic model of residential real estate prices discussed by Anglin and Gencay (1996). Their baseline model is a linear regression of the logarithm of sales prices on an intercept and 11 attributes. The attributes are indicated in the first column of Table 1. All variables beginning with “#” are positive integers, log(Lot size) is continuous, and all other variables are dichotomous (1 if present and 0 if not). The data consist of 546 transactions during July, August, and September 1987 in metropolitan Windsor, Ontario. The least squares estimates match those reported in Anglin and Gencay (1996).

The form of the prior distribution is the one discussed in Section 2.1 for the normal linear model. The normal distribution for the coefficient vector has mean $0$, the precision matrix is diagonal, and standard deviations are chosen to allow reasonable values of the coefficients. The prior distribution of the precision parameter is $0.12h \sim \chi^2(3)$ so that $h$ has prior mean 25 and standard deviation about 20. The posterior simulator is the Gibbs sampling algorithm described in Section 3.3, based on the conditional distributions Eqs. (17) and (18). The posterior simulation structure was created using the BACC commands `minst` and `postsim`. Creation of 10,000 records in the posterior simulation structure required

roughly 8 sec.* The last four columns of Table 1 report results from `expect1` using this posterior simulation structure, discarding the first 1000 iterations. Initial values here, and in all other examples discussed in this chapter, were drawn from the prior distribution, and in every case the first 1000 draws were discarded. The `expect1` computations took a quarter of a second. Posterior means and standard deviations are close to the least squares values, reflecting the lack of information in the prior relative to the data set. The numerical standard error (NSE) of each posterior mean is given for \( L = 0.08(M - r) \) as described in the previous section 3.7.† The NSEs imply accuracy of more than two figures past the decimal in the posterior means, and the relative numerical efficiency (RNE) indicates that numerical accuracy is comparable to what would have been achieved with i.i.d. drawings directly from the posterior distribution.

The second example is a probit model of women’s labor force participation, based on the one presented in Geweke and Keane (1999). The data consist of 1555 observations of women in the 1987 Panel Survey of Income Dynamics. The choice variable is 1 if a woman reports positive hours of work for 1987. The covariates are indicated in the first

<table>
<thead>
<tr>
<th>Coefficient (RNE)</th>
<th>Ordinary least squares</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>s.e.</td>
<td>Mean</td>
</tr>
<tr>
<td>Intercep</td>
<td>7.745</td>
<td>0.216</td>
<td>0</td>
</tr>
<tr>
<td>Driveway</td>
<td>0.110</td>
<td>0.028</td>
<td>0</td>
</tr>
<tr>
<td>Recreation room</td>
<td>0.058</td>
<td>0.026</td>
<td>0</td>
</tr>
<tr>
<td>Finished basement</td>
<td>0.104</td>
<td>0.021</td>
<td>0</td>
</tr>
<tr>
<td>Gas hot water</td>
<td>0.179</td>
<td>0.043</td>
<td>0</td>
</tr>
<tr>
<td>Central air</td>
<td>0.166</td>
<td>0.021</td>
<td>0</td>
</tr>
<tr>
<td>#Garage stalls</td>
<td>0.048</td>
<td>0.011</td>
<td>0</td>
</tr>
<tr>
<td>Good nbhd</td>
<td>0.132</td>
<td>0.023</td>
<td>0</td>
</tr>
<tr>
<td>log(Lot size)</td>
<td>0.303</td>
<td>0.027</td>
<td>0</td>
</tr>
<tr>
<td>#Bedrooms</td>
<td>0.034</td>
<td>0.014</td>
<td>0</td>
</tr>
<tr>
<td>#Full bathrooms</td>
<td>0.166</td>
<td>0.020</td>
<td>0</td>
</tr>
<tr>
<td>#Stories</td>
<td>0.092</td>
<td>0.013</td>
<td>0</td>
</tr>
</tbody>
</table>

*All execution times are given for a Pentium II processor with 64MB of RAM.
†This value of \( L \) is also used to report NSE and RNE in all other examples in this chapter’s. Results using other values of \( L \) are similar.
column of Table 2. “Black,” “Married,” and “Kids” are dichotomous. Age is measured in years and interacted with “Married” and its negation. “Education” is years of completed schooling. “Spouse$” is husband’s income and “Family$” is unearned household income, in dollars for the year 1987. Work experience is measured in cumulative hours since the woman became a household head or spouse. For each unmarried woman with children, AFDC is the monthly cash support she would receive if she did not work. “Food$” is the monthly food-stamp allotment to which a woman’s household would be entitled if she did not work. The last two variables differ according to the state of residence. The prior distributions of the coefficients are independent normal, each with mean zero and standard deviation chosen to permit large but reasonable values to be within two standard deviations of zero. Details of sample screening, variable descriptions, and prior construction are given in Geweke and Keane (1999).

Conventional maximum likelihood estimates, the posterior mode, and approximate posterior standard deviations based in the usual way on the Hessian of the log-posterior at the mode* are presented in Table 2. It

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Maximum likelihood</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mode</td>
<td>Asymptotic s.e.</td>
<td>Prior Mean</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.22</td>
<td>0.520</td>
<td>0</td>
</tr>
<tr>
<td>Black</td>
<td>0.109</td>
<td>0.105</td>
<td>0</td>
</tr>
<tr>
<td>Age-Single</td>
<td>-0.0611</td>
<td>0.0132</td>
<td>0</td>
</tr>
<tr>
<td>Age-Married</td>
<td>-0.0874</td>
<td>0.0128</td>
<td>0</td>
</tr>
<tr>
<td>Education</td>
<td>0.113</td>
<td>0.0274</td>
<td>0</td>
</tr>
<tr>
<td>Married</td>
<td>0.682</td>
<td>0.522</td>
<td>0</td>
</tr>
<tr>
<td>Kids</td>
<td>-0.488</td>
<td>0.171</td>
<td>0</td>
</tr>
<tr>
<td>#Kids</td>
<td>-0.0506</td>
<td>0.0652</td>
<td>0</td>
</tr>
<tr>
<td>Spouse$</td>
<td>$-1.78 \times 10^{-5}$</td>
<td>$3.46 \times 10^{-6}$</td>
<td>0</td>
</tr>
<tr>
<td>Family$</td>
<td>$1.86 \times 10^{-6}$</td>
<td>$7.25 \times 10^{-6}$</td>
<td>0</td>
</tr>
<tr>
<td>AFDC</td>
<td>$-5.02 \times 10^{-4}$</td>
<td>$3.85 \times 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>Food$</td>
<td>$-0.00211$</td>
<td>$6.94 \times 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>WorkExp</td>
<td>$1.36 \times 10^{-4}$</td>
<td>$9.50 \times 10^{-6}$</td>
<td>0</td>
</tr>
</tbody>
</table>

*For details and the asymptotic justification for this approximation see Bernardo and Smith (1994, Sect. 5.3) and the references cited therein.
is clear from Table 2 that the prior distribution is informative, relative to the data: prior standard deviations of 6 of the 13 coefficients are less than the corresponding maximum likelihood asymptotic standard errors. Posterior standard deviations are less than asymptotic standard errors in every case, and for several coefficients the posterior mode is closer to the prior mode than to the likelihood mode.

The Gibbs sampler for the probit model, developed in Albert and Chib (1993) and described in Section 3.3, is the posterior simulation algorithm used to construct the posterior simulation structure. As with the first example, this structure is created using the BACC commands minst and postsim. Creation of a posterior simulation structure with 10,000 records required 410 sec. Posterior means, standard deviations, and measures of numerical accuracy, based on the last 9000 records, are presented in Table 3. The posterior moments are quite close to the approximation at the posterior mode, very likely reflecting a posterior distribution that is close to multivariate normal. The average relative numerical efficiency for the coefficients indicates that the same accuracy could have been achieved with about half the number of iterations, had an i.i.d. sample been drawn directly from the posterior distribution.

### 4 MODEL COMPARISON

Section 2.3 demonstrated that given prior probabilities over models, and prior probability distributions for parameter vectors within models, there is a complete theory of model combination and model comparison. The

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Mean</th>
<th>Stand. dev.</th>
<th>NSE</th>
<th>RNE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.2135</td>
<td>0.176</td>
<td>0.00376</td>
<td>0.245</td>
</tr>
<tr>
<td>Black</td>
<td>0.01538</td>
<td>0.0772</td>
<td>0.00173</td>
<td>0.222</td>
</tr>
<tr>
<td>Age-Single</td>
<td>−0.010258</td>
<td>0.00381</td>
<td>4.65 × 10⁻⁵</td>
<td>0.745</td>
</tr>
<tr>
<td>Age-Married</td>
<td>−0.028012</td>
<td>0.00650</td>
<td>1.08 × 10⁻⁴</td>
<td>0.406</td>
</tr>
<tr>
<td>Education</td>
<td>0.0022905</td>
<td>0.00411</td>
<td>4.62 × 10⁻⁵</td>
<td>0.876</td>
</tr>
<tr>
<td>Married</td>
<td>0.17867</td>
<td>0.118</td>
<td>0.00122</td>
<td>1.038</td>
</tr>
<tr>
<td>Kids</td>
<td>−0.36476</td>
<td>0.132</td>
<td>0.00262</td>
<td>0.279</td>
</tr>
<tr>
<td>#Kids</td>
<td>−0.15209</td>
<td>0.0438</td>
<td>7.30 × 10⁻⁴</td>
<td>0.400</td>
</tr>
<tr>
<td>Spouse$</td>
<td>−7.2318 × 10⁻⁶</td>
<td>2.25 × 10⁻⁶</td>
<td>3.62 × 10⁻⁸</td>
<td>0.432</td>
</tr>
<tr>
<td>Family$</td>
<td>−1.1449 × 10⁻⁶</td>
<td>2.91 × 10⁻⁶</td>
<td>3.81 × 10⁻⁸</td>
<td>0.648</td>
</tr>
<tr>
<td>AFDC</td>
<td>−5.7610 × 10⁻⁴</td>
<td>3.03 × 10⁻⁴</td>
<td>4.25 × 10⁻⁶</td>
<td>0.565</td>
</tr>
<tr>
<td>Food$</td>
<td>−0.0012176</td>
<td>4.33 × 10⁻⁴</td>
<td>5.91 × 10⁻⁶</td>
<td>0.597</td>
</tr>
<tr>
<td>WorkExp</td>
<td>1.1672 × 10⁻⁴</td>
<td>8.46 × 10⁻⁶</td>
<td>4.24 × 10⁻⁷</td>
<td>0.044</td>
</tr>
</tbody>
</table>
central technical task in implementing the theory is calculation of the marginal likelihood \( p(Y_T | A) = \int_q p(Y_T | A, \theta) p(\theta | A) d\theta \). The marginal likelihood cannot, in general, be cast in the form of a posterior moment, Eq. (5), and therefore the posterior simulation methods of Section 3, which have proven useful in obtaining posterior moments in a single model, are not directly applicable to this problem. A decade ago, there were essentially no methods developed for the numerical approximation of marginal likelihoods or Bayes factors and results were limited to a handful of cases for which there were analytical results or asymptotic approximations. Now, it is possible to attain good and generic approximations to marginal likelihoods in most cases; however, some models with large numbers of latent variables remain troublesome.

This section provides several approaches to the approximation of marginal likelihoods, with an emphasis on generic methods that are consistent as the number of simulations increases. Generic methods exclude those that are ingenious but specific to particular situations as well as methods that rely on asymptotic approximations rather than simulation. Many of these methods are discussed in a comprehensive review article by Kass and Raftery (1996). We discuss here a method that works well with importance sampling and the Hastings–Metropolis algorithm; a method specific to the Gibbs sampler; and a generic method that works well with most posterior simulators regardless of the algorithm employed. For the last method, we describe publicly available software designed to work with the most commonly used computing platforms in econometrics. Some examples illustrate the numerical accuracy that can be attained, and provide comparisons of some of the different methods of approximating marginal likelihoods.

### 4.1 Importance Sampling

Suppose that \( j(\theta) \), with support \( \Theta \), is the probability density function (not just a kernel) with respect to the measure \( d\theta \) of an importance sampling distribution for the posterior density \( p(\theta | Y_T, A) \propto p(\theta | A) p(Y_T | \theta, A) \), where \( p(\theta | A) \) is the properly normalized prior density and \( p(Y_T | \theta, A) \) is the properly normalized data density. Define the weighting function \( w(\theta) = p(\theta | A) p(Y_T | \theta, A) / j(\theta) \).

**Corollary 3**

Let \( j(\theta) \) be the importance sampling density in an importance sampling algorithm. Suppose the support of \( j(\theta) \) includes \( \Theta \). Then
\[ \tilde{w}_M = M^{-1} \sum_{m=1}^{M} w(\theta^{(m)}) \stackrel{a.s.}{\longrightarrow} \int_{\Theta} w(\theta) \, d\nu(\theta) \]

\[ = \int_{\Theta} p(\theta|A) p(Y_T|\theta, A) \, d\nu(\theta) = p(Y_T|A) \]

If \( w(\theta) \) is bounded above then

\[ M^{1/2}[\tilde{w}_M - p(Y_T|A)] \xrightarrow{d} N(0, \sigma^2) \text{ and} \]

\[ M^{-1} \sum_{m=1}^{M} \left[ w(\theta^{(m)}) - \tilde{w}_M \right]^2 \xrightarrow{a.s.} \sigma^2 \]

**Proof**

Immediate from Theorem 2 and Corollary 1.

The first application of this idea is that of Geweke (1989a); see also Gelfand and Dey (1994) and Raftery (1995). Since \( w(\theta^{(m)}) \) must be computed from each iteration of the importance sampling algorithm in any event and the normalizing constant for \( j(\theta) \) is usually known, this simulation-consistent approximation of \( p(Y_T|A) \) may be obtained at essentially no additional cost.

### 4.2 The Gibbs Sampler

In the case of the Gibbs sampler there is a different procedure due to Chib (1995) that provides accurate evaluations of the marginal likelihood, at the cost of additional simulations. Suppose that the output from the blocking \( \theta' = (\theta'_1, \ldots, \theta'_B) \) is available, and that the conditional p.d.f.'s \( p(\theta_j|\theta_{\neq j}, Y_T, A) \) can be evaluated in closed form for all \( j \). (This latter requirement is generally satisfied.)

From Eqs. (1) and (2):

\[ p(Y_T|A) = p(\tilde{\theta}|A) p(Y_T|\tilde{\theta}, A) / p(\tilde{\theta}|Y_T, A) \]  \hspace{1cm} (50)

for any \( \tilde{\theta} \in \Theta \). Typically \( p(Y_T|\tilde{\theta}, A) \) and \( p(\tilde{\theta}|A) \) can be evaluated in closed form, but \( p(\tilde{\theta}|Y_T, A) \) cannot. A marginal/conditional decomposition of \( p(\tilde{\theta}|Y_T, A) \) is

\[ p(\tilde{\theta}|Y_T, A) = p(\tilde{\theta}_{(1)}|Y_T, A) p(\tilde{\theta}_{(2)}|Y_T, \tilde{\theta}_{(1)}, A) \cdots \]

\[ \times p(\tilde{\theta}_{(B-1)}|Y_T, \tilde{\theta}_{(1)}, \ldots, \tilde{\theta}_{(B-2)}, A) \]  \hspace{1cm} (51)
The first term in the product of $B$ terms can be approximated from the output of the posterior simulator because

$$M^{-1} \sum_{m=1}^{M} p(\tilde{\theta}_{(1)}|Y_T, \tilde{\theta}_{(2)^{1}}, \ldots, \tilde{\theta}_{(B)^{M}}, A) \xrightarrow{a.s.} p(\tilde{\theta}_{(1)}|Y_T, A)$$

To approximate $p(\tilde{\theta}_{(b)}|Y_T, \tilde{\theta}_{(1)}, \ldots, \tilde{\theta}_{(b-1)}, A)$, first execute the Gibbs sampler with the parameters in the first $b - 1$ blocks fixed at the indicated values, thus producing a sequence $\{\tilde{\theta}_{(b)^{1}}, \ldots, \tilde{\theta}_{(b)^{M}}\}$ from the conditional posterior. Then

$$M^{-1} \sum_{m=1}^{M} p(\tilde{\theta}_{b}|Y_T, \tilde{\theta}_{(1)}, \ldots, \tilde{\theta}_{(b-1)}, \tilde{\theta}_{(b)^{1}, (b)^{1}}, \ldots, \tilde{\theta}_{(b)^{M}, (b)}, A) \xrightarrow{a.s.} p(\tilde{\theta}_{(b)}|Y_T, \tilde{\theta}_{(1)}, \ldots, \tilde{\theta}_{(b-1)}, A)$$

These approximations are then used in Eqs. (50) and (51) to obtain the approximation to the marginal likelihood. In general, this method is more efficient the greater is $p(\tilde{\theta}|Y_T, A)$, so in many applications it is natural to choose $\tilde{\theta}$ near the posterior mode. It is straightforward to apply the methods of Section 3.7 to evaluate the numerical accuracy of the final approximation to the marginal likelihood, using standard delta methods. See Chib (1995) on these and other important practical details.

### 4.3 Modified Harmonic Mean

Gelfand and Dey (1994) observe that for any p.d.f. $f(\theta)$, whose support is contained in $\Theta$,

$$E\left[\frac{f(\theta)}{p(\theta|A)p(Y_T|\theta, A)} \bigg| Y_T, A\right]$$

$$= \int_{\Theta} \frac{f(\theta)}{p(\theta|A)p(Y_T|\theta, A)} p(\theta|Y_T, A) \, dv(\theta)$$

$$= \int_{\Theta} \frac{f(\theta)}{p(\theta|A)p(Y_T|\theta, A)} \cdot \frac{p(\theta|A)p(Y_T|\theta, A)}{\int_{\Theta} p(\theta|A)p(Y_T|\theta, A) \, dv(\theta)} \, dv(\theta)$$

$$= \int_{\Theta} \frac{f(\theta)}{p(\theta|A)p(Y_T|\theta, A)} \, dv(\theta) \int_{\Theta} p(\theta|A)p(Y_T|\theta, A) \, dv(\theta)$$

$$= \int_{\Theta} \frac{f(\theta)}{p(\theta|A)p(Y_T|\theta, A)} \, dv(\theta)$$

$$= p(Y_T|A)^{-1}$$

(52)

Thus, the posterior mean of the function of interest $f(\theta)/p(\theta|A)p(Y_T|\theta, A)$ is $p(Y_T|A)^{-1}$. It is, therefore, a candidate for approximation by a
posterior simulator. If \( f(\theta)/p(\theta|A)p(Y_T|\theta, A) \) is bounded above, then the approximation is simulation consistent and the rate of convergence is likely to be practical.

It is not difficult to guarantee the boundedness condition in Eq. (52). Consider first the case in which \( \Theta = \mathbb{R}^k \). From the output of the posterior simulator define

\[
\hat{\theta}_M = \frac{\sum_{m=1}^M w(\theta^{(m)}) \theta^{(m)}}{\sum_{m=1}^M w(\theta^{(m)})}
\]

and

\[
\hat{\Sigma}_M = \frac{\sum_{m=1}^M w(\theta^{(m)})(\theta^{(m)} - \hat{\theta}_M)(\theta^{(m)} - \hat{\theta}_M)'}{\sum_{m=1}^M w(\theta^{(m)})}
\]

(It is not essential that the posterior mean and variance of \( \theta \) exist.) Then, for some \( p \in (0, 1) \), define \( \hat{\Theta}_M = \{ \theta : (\theta - \hat{\theta}_M)' \hat{\Sigma}_M^{-1} (\theta - \hat{\theta}_M) \leq \chi^2_{p}(k) \} \) and take

\[
f(\theta) = p^{-1}(2\pi)^{-k/2}|\hat{\Sigma}_M|^{-1/2}\exp\left[-0.5(\theta - \hat{\theta}_M)' \hat{\Sigma}_M^{-1} (\theta - \hat{\theta}_M)\right] \chi_0(\hat{\theta}_M, \theta)
\]  (53)

If the posterior density is uniformly bounded away from 0 on every compact subset of \( \Theta \), then the function \( f(\theta)/p(\theta|A)p(Y_T|\theta, A) \) possesses posterior moments of all orders. For a wide range of regular problems, this function will be approximately constant on \( \hat{\Theta}_M \), which is nearly ideal. In most situations smaller values of \( p \) will result in better behavior of \( f(\theta)/p(\theta|A)p(Y_T|\theta, A) \) over the domain \( \hat{\Theta}_M \), but greater simulation error due to a smaller number of \( \theta^{(m)} \in \hat{\Theta}_M \); there is almost no incremental cost in carrying out the computations for several values of \( p \) rather than a single value of \( p \).

So long as \( \hat{\Theta}_M \subseteq \Theta \), \( \int_{\hat{\Theta}_M} f(\theta) \, d\theta = 1 \). If not, the domain of integration must be redefined to be \( \hat{\Theta}_M \cap \Theta \). In this case a new normalizing constant for \( f(\theta) \) can be well approximated by taking a sequence of i.i.d. draws \( \{\theta^{(t)}\} \) from the original distribution Eq. (53) with domain \( \hat{\Theta}_M \), and then averaging \( \chi_0(\hat{\theta}_M, \theta) \).

Frequently the behavior of \( f(\theta)/p(\theta|A)p(Y_T|\theta, A) \) can be improved by reparameterization of \( \theta \) to \( \zeta = h(\theta) \), where \( h \) is a one-to-one function. Of course, the prior density must then be adjusted by the Jacobian of transformation. If the support \( Z \) of \( p(\zeta|A) \) is \( \mathbb{R}^k \), then \( \int_Z f(\zeta) \, d\zeta = 1 \) for \( f \) constructed as indicated in Eq. (53). For example, if this method is used to

*The weighting function \( w(\theta) \) is defined in Theorem 2 in the case of importance sampling. For MCMC algorithms \( w(\theta) = 1 \).
approximate the marginal likelihood in the standard linear model, Eqs. (7)–(9), transformation of \( h \) to \( \log(h) \) guarantees the support condition, and generally results in more accurate approximation of \( p(Y_T|A) \).

The numerical accuracy of the approximation can be evaluated using the methods of Section 3.7, as detailed in Section 4.5.

### 4.4 Improving Numerical Approximations

In many instances a portion of the marginal likelihood \( P(Y_T|A) = \int p(\theta|A)p(Y_T|\theta, A)\,d\theta \) may be evaluated analytically. Suppose that

\[
\int p(\theta|A)p(Y_T|\theta, A)\,d\theta = \int \int p(\theta_1, \theta_2|A)p(Y_T|\theta_1, \theta_2, A)\,d\theta_1\,d\theta_2 = \int r(\theta_1)\,d\theta_1
\]

where \( r(\theta_1) = \int p(\theta_1, \theta_2|A)p(Y_T|\theta_1, \theta_2, A)\,d\theta_2 \) can be evaluated analytically. The modified harmonic mean method can be applied directly to the simulated values \( \theta^{(m)} \), using \( r(\theta^{(m)}) \) in lieu of \( p(\theta^{(m)}|A)p(Y_T|\theta^{(m)}, A) \) and tailoring \( f(\theta) \) to \( r(\theta) \) rather than to \( p(\theta|A)p(Y_T|\theta, A) \). Similar adjustments can be made for importance sampling. Because the dimension of integration is lower, the resulting approximation will typically be more accurate. In the case of the method employing the Gibbs sampler described in Section 4.2 this preliminary evaluation will eliminate at least one of the blocks for which the auxiliary simulations must be undertaken.

An example of this procedure is provided by earlier results for the standard linear model. The entire posterior kernel in standard form is Eq. (13), but in Section 2.3 we found the marginal likelihood conditional on \( h \), Eq. (25). The latter expression is a function of a single unknown parameter, whereas the former is a function of \( k + 1 \) unknown parameters.

The probit model described in Section 2.4 provides a second example. In this case there are \( T + k \) unknown parameters (\( T \) latent variables and \( k \) coefficients). The modified harmonic mean method in this case is completely unwieldy, since it would require storing a very large amount of posterior simulator output, and generation of the requisite \( T + k \) truncated normal random variables would require the factorization of a matrix of the same order. In this case, integration of the \( T \) latent variables is straightforward, and leads to the product of the prior density for the coefficients and the likelihood function as typically written:
More generally, in models with latent variables, accurate evaluation of the marginal likelihood requires that it be possible to perform the integration over the space of such variables analytically.

### 4.5 Software

The `mlike` command computes estimates of the log marginal likelihood using the modified harmonic mean posterior simulation method, given a model instance for which the user has generated a posterior simulation structure using the `postsim` command. The user optionally supplies a vector of values of the p parameter of Section 4.3. Otherwise the single default value of 0.5 applies. For each value of p, the `mlike` command computes

\[
(2\pi)^{-k/2} |H_p|^{1/2} \exp\left[ -\frac{1}{2} (\beta - \beta_p)^T H_p (\beta - \beta_p) \right] \\
\times \prod_{t=1}^{T} \{ d_t \Phi(\beta' x_t) + (1 - d_t) [1 - \Phi(\beta' x_t)] \}
\]

and reports minus the logarithm of this value. Also, for each value of p, it optionally provides numerical standard errors for each value of L (described in Section 3.7) supplied by the user. The numerical standard error for the log marginal likelihood is obtained as the standard error of the quantity in Eq. (54) divided by that quantity. This latter standard error is computed as described in Section 3.7.

Parameter transformations to improve the behavior of \( f(\theta)/p(\theta|A) \) \( p(Y_T|\theta) \), and simulations to handle cases where \( \Theta_M \not\subset \Theta \), discussed in Section 4.3, are performed automatically by the `mlike` command. The parameter transformations and the indicator function \( \chi_\Theta \) form part of the definition of a model, supplied by the model developer. They are invisible to the user.

### 4.6 Examples

For the regression model described in Section 3.8, the marginal likelihood was approximated using the BACC command `mlike`. The software incorporates a reparameterization of the precision from \( h \) to \( \log h \). After this reparameterization \( \Theta_A = \Theta^h \), the `mlike` execution time is under half a second for this case.

The top panel of Table 4.1 provides results using \( p = 0.9, 0.5, \) and 0.1 in Eq. (53). Computation with \( p = 0.9 \) provides the most accurate
assessment. In view of the good approximation of the posterior by a multivariate normal distribution, it is not surprising that a more inclusive \( f(\theta) \) yields more accurate results. Differences in approximations for different values of \( p \) are consistent with the numerical standard errors (NSEs).

To illustrate the use of the approximated marginal likelihood in the construction of Bayes factors, two variants on the model set forth in Section 3.8 were constructed by making two changes in the prior distribution. In the first change the mean of the prior distributions of all slope coefficients is shifted to the value of the standard deviation which in turn is unchanged: i.e., the prior distribution is changed from \( N(0, 0.1^2) \) to \( N(1, 0.1^2) \) for all covariates except \( \log(\text{Lot size}) \), for which the prior distribution is shifted from \( N(0, 0.3^2) \) to \( N(0.3, 0.3^2) \). Log marginal likelihood approximations for this model are given in the middle panel of Table 4. Finally, the standard deviations in these priors are reduced by

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Regression Model: Marginal Likelihoods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log marginal likelihood</td>
</tr>
<tr>
<td>First prior (zero center)</td>
<td></td>
</tr>
<tr>
<td>( p = 0.9 )</td>
<td>46.081</td>
</tr>
<tr>
<td>( p = 0.5 )</td>
<td>46.082</td>
</tr>
<tr>
<td>( p = 0.1 )</td>
<td>46.093</td>
</tr>
<tr>
<td>Second prior (nonzero center)</td>
<td></td>
</tr>
<tr>
<td>( p = 0.9 )</td>
<td>52.148</td>
</tr>
<tr>
<td>( p = 0.5 )</td>
<td>52.151</td>
</tr>
<tr>
<td>( p = 0.1 )</td>
<td>52.162</td>
</tr>
<tr>
<td>Third prior (nonzero center, higher precision)</td>
<td></td>
</tr>
<tr>
<td>( p = 0.9 )</td>
<td>56.365</td>
</tr>
<tr>
<td>( p = 0.5 )</td>
<td>56.369</td>
</tr>
<tr>
<td>( p = 0.1 )</td>
<td>56.383</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Probit Model: Marginal Likelihoods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log marginal likelihood</td>
</tr>
<tr>
<td>( p = 0.9 )</td>
<td>-564.71</td>
</tr>
<tr>
<td>( p = 0.5 )</td>
<td>-564.70</td>
</tr>
<tr>
<td>( p = 0.1 )</td>
<td>-564.69</td>
</tr>
</tbody>
</table>
half: now all priors are $N(0,1,0.05^2)$ except for $\log(\text{Lot size})$, which is $N(0.3,0.15^2)$. Log marginal likelihoods for this model are shown in the lower panel of Table 4.

Using the approximation based on $p = 0.9$, the log Bayes factor in favor of the last model, versus the first, is approximately 10.285 and the associated NSE is 0.006. Thus, the Bayes factor is almost certainly (based on NSE×3) between 28,767 and 29,822.

In the probit model example, the marginal likelihood can be approximated using the modified harmonic mean method implemented in mlk. The evaluation of the likelihood function for the probit model as discussed at the end of Section 4.4 is used. The results are shown in Table 5. For the same reasons as in the regression model the approximation is quite accurate and is better for larger values of $p$.

5 BAYESIAN COMMUNICATION

For a subjective Bayesian decision maker the computation of the posterior moments $E[h(\omega)|\mathbf{Y}_T,A]$ for suitable models, priors, and functions of interest is typically the final objective of inference. For an investigator reporting results for other potential decision makers, however, the situation is different. In the language of Hildreth (1963) these decision makers are remote clients, whose priors and functions of interest are not known to the investigator.

What should the investigator report? Traditionally, published papers report a few posterior moments and more rarely some indication of sensitivity to prior distributions, and alternative data densities may be given. Such information is generally much too limited. At the other extreme, the investigator may simply report some likelihood functions, but this leaves most of the work to the client. Investigators almost never report marginal likelihoods, thereby leaving unrealized the promise inherent in model averaging.

5.1 Posterior Reweighting

An investigator will have carried through formal inference for a set of models $A_1,\ldots,A_J$. This collection will reflect the process of model development, and a public report of the investigator’s work should at least summarize this process. In the ideal situation described by Poirier (1988), clients have agreed to disagree in terms of the prior. Since the set of models that exists in any meaningful sense is the set publicly reported,
collectively investigators will have provided the grand model in which variation of the prior is the basis of formal discourse in normal science.*

Corresponding to each model \( A_1, \ldots, A_J \) included in an investigation, there is a posterior simulation structure of the form described in Section 3.7. It is a simple matter to save these simulation structures and make them available at an ftp or web site, and for any client to obtain them for the purpose of the manipulations described here.

Given the posterior simulation structure a client can immediately compute numerical approximations to posterior moments not reported or even considered by the investigator. Specifically, suppose a client wishes to know

\[
E[h(\omega)|Y_T, A]
\]

where \( \omega \sim p(\omega|Y_T, A) \) is specified by the client and \( p(Y_T|\theta, A) \) and \( p(\theta|A) \) have been specified by the investigator. Corresponding to each \( \theta^{(m)} \) reported in the posterior simulation structure the client forms

\[
g(Y_T, \theta^{(m)})
\]

with the property

\[
E[g(Y_T, \theta)|Y_T, \theta, A] = E[h(\omega)|Y_T, \theta, A],
\]

and then computes

\[
\bar{g}_M = \sum_{m=1}^{M} w(\theta^{(m)}) g(\theta^{(m)})/ \sum_{m=1}^{M} w(\theta^{(m)}).
\]

If the investigator’s posterior simulator is ergodic then

\[
\bar{g}_M \xrightarrow{a.s.} \bar{g} = E[h(\omega)|Y_T, A]
\]

and if it is uniformly ergodic then

\[
M^{1/2}(\bar{g}_M - \bar{g}) \xrightarrow{a.s.} N(0, \sigma^2).
\]

For simple functions \( h(\omega) \) computation of \( \bar{g}_M \) amounts to spreadsheet arithmetic. More elaborate functions of interest may involve simulations, but in all cases these computations are precisely those which economists undertake as a routine matter when investigating the implications of a model.

For example, a client reading a research report might be skeptical that the investigator’s model, prior, and data set provide much information about the effects of an interesting change in a policy variable on the outcome in question. If the simulator output matrix is available electronically the client can obtain the exact (up to numerical approximation error, which can also be evaluated) answer to his query, without arising from his office chair, in considerably less time than required to read the research report.

The social contribution of the investigator in this context is clear. She enables clients to incorporate the effects of uncertainty about parameters in a specified model consisting of \( p(Y_T|\theta, A) \) and \( p(\theta|A) \), in reaching conclusions or decisions of the client’s choosing, which can be addressed by the model. This contribution extends in an obvious way to uncertainty about models, so long as a posterior simulation structure has been provided by an investigator for each model considered.

With a small amount of additional effort the client can modify many of the investigator’s assumptions. Suppose the client wishes to evaluate

\[
E[h(\omega)|Y_T, A^*],
\]

where the model \( A^* \) differs from the model \( A \) only in the

*The term normal science is used here as in Kuhn (1970).
specification of the prior distribution \( p(\theta|A^*) \neq p(\theta|A) \); i.e., \( p(Y_T|\theta, A) = p(Y_T|\theta, A^*) \). Suppose further that the support of the investigator’s prior distribution includes the support of the client’s prior. The investigator’s posterior density may be then regarded as an importance sampling density for the client’s prior density. The client reweights the investigator’s \( \{\theta^{(m)}\}_{m=1}^{M} \) using the function:

\[
w(\theta; A^*) = \frac{p(\theta|Y_T, A^*)}{p(\theta|Y_T, A)} \times \frac{p(\theta|A^*)p(Y_T|\theta, A^*)}{p(\theta|A)p(Y_T|\theta, A)} = \frac{p(\theta|A^*)}{p(\theta|A)}
\]

The client then approximates his posterior moment \( E[g(\theta)|Y_T, A^*] \) by

\[
\bar{g}_M = \frac{\sum_{m=1}^{M} w(\theta^{(m)}; A^*) w(\theta^{(m)}) g(Y_T \theta^{(m)})}{\sum_{m=1}^{M} w(\theta^{(m)}; A^*) w(\theta^{(m)})}
\]

Geweke (2000) shows that if \( \{\theta^{(m)}\} \) is ergodic with invariant distribution \( p(\theta|Y_T, A) \) with \( \theta \), \( E[g(Y_T, \theta)|Y_T, A^*] \) exists and is finite, and the support of \( p(\theta|A) \) includes the support of \( p(\theta|A^*) \), then \( \bar{g}_M \xrightarrow{a.s.} E[g(Y_T, \theta)|Y_T, A^*] = \bar{g}^* \); and if in addition \( \{\theta^{(m)}\} \) is uniformly ergodic, var[\( g(Y_T, \theta)|Y_T, A \)] exists and is finite, and \( p(\theta|A^*)/p(\theta|A) \) is bounded above, then there exists \( \sigma^2 > 0 \) such that \( M^{1/2}(\bar{g}_M - \bar{g}^*) \xrightarrow{d} N(0, \sigma^2) \).

Efficiency of the reweighting scheme requires some similarity of \( p(\theta|A^*) \) and \( p(\theta|A) \), as illustrated subsequently in Section 5.3. In particular, both reasonable convergence rates and the use of a central limit theorem to assess numerical accuracy essentially require that \( p(\theta|A^*)/p(\theta|A) \) be bounded. Across a set of diverse clients this condition is more likely to be satisfied the more diffuse is \( p(\theta|A) \), and is trivially satisfied for the (possibly improper) prior \( p(\theta|A) \propto \chi_0(\theta|A) \) if the client’s prior is bounded. In the latter case the reweighting scheme will be efficient so long as the client’s prior is uninformative relative to the likelihood function. This condition is stated precisely in Theorem 2 of Geweke (1989b). Relative numerical efficiency will indicate situations in which the reweighting scheme is inefficient. If the investigator chooses to use an improper prior for reporting, it is of course incumbent on her to verify the existence of the posterior distribution and convergence of her posterior simulator.

Including \( p(\theta^{(m)}|A) \) in the standard posterior simulation structure avoids the need for every client who wishes to impose his own priors to re-evaluate the investigator’s prior. Of course, \( p(\theta|A^*) \) need not be the client’s subjective prior: it may simply be a device by which the client, functioning as another investigator, explores the robustness of results with respect to alternative reasonable priors.
The reweighting scheme permits some updating of the investigator’s results at relatively low cost. If observations \( T + 1, \ldots, T + f \) beyond the \( T \) originally used have become available, then

\[
p(\theta | Y_{T+f}, A) \propto p(\theta | A) p(\theta) = p(\theta | A) p(\theta) \prod_{s=T+1}^{T+f} p(y_s | Y_{s-1}, \theta, A)
\]

\[
\propto p(\theta | Y_{T}, A) \prod_{s=T+1}^{T+f} p(y_s | Y_{s-1}, \theta, A)
\]

The client therefore forms the approximation to the updated posterior moment \( \mathbb{E}[g(Y_{T+f}) | Y_{T+f}, A] \):

\[
\mathbb{E}_{M}^{*} = \frac{\sum_{m=1}^{M} w(\theta^{(m)}; Y_{T+f}) w(\theta^{(m)}) g(Y_{T+f}, \theta^{(m)})}{\sum_{m=1}^{M} w(\theta^{(m)}; Y_{T+f}) w(\theta^{(m)})} \xrightarrow{a.s.} \mathbb{E}[g(Y_{T+f}) | Y_{T+f}, A] = \mathbb{E}^{*}
\]

with \( w(\theta; Y_{T+f}) = \prod_{s=T+1}^{T+f} p(y_s | Y_{s-1}, \theta, A) \). If \( f \) is small relative to \( T \), and there is no major change in the data generating process between \( T \) and \( T + f \), the new approximation will be efficient. However, as \( f \) grows, efficiency diminishes and at some point the approximation \( \mathbb{E}^{*} \) becomes too inaccurate to be useful. This process also requires evaluation of the likelihood function, which usually involves more technical difficulties than evaluation of priors or functions of interest.

5.2 Software

No special BACC commands are required for reweighting or computing new functions of interest. The general-purpose commands of the mathematical application itself are sufficient. The user simply calculates new functions of interest and/or computes new weights and passes them as input to the expect1 command.

5.3 Examples

Let us return to the regression model of hedonic pricing introduced in Section 3.8. Suppose that an investigator wishes to provide a posterior simulation structure with the intention that clients will impose their own priors by reweighting the output. To this end the investigator should use a prior distribution that is uninformative relative to the data. Illustrating what such an investigator might do, prior distributions \( N(0, 1^2) \) for all slope coefficients, except \( N(0, 3^2) \) for log(Lot size), and \( 0.04h \sim \chi^2(1) \) for precision, were chosen and the corresponding posterior simulation structure with 10,000 replications was created. To mimic a client, the
tightest of the three prior distributions described in Section 4.6 was chosen: \( N(0.1, 0.05^2) \) for all slope coefficients except \( N(0.3, 0.15^2) \) for lot size, and \( 0.12h \sim \chi^2(3) \) for precision. A new vector of log weights was created by taking the log of the client’s prior minus the log of the investigator’s prior. Then \texttt{expect1} was used with this new vector of log weights to obtain posterior moments for the coefficients; this required a quarter of a second.

The results of this exercise are displayed in Table 6. The left panel provides the results that would have been obtained had the client directly executed the posterior simulator corresponding to his prior. The accuracy of the numerical approximation of the posterior moments is similar to that exhibited for the less informative prior in Table 1. The right panel displays the results the client obtains by reweighting the investigator’s simulator output. The chi-squared statistics from \texttt{expectN} (last column) comparing the posterior means approximated in these two different ways indicate no difficulties with the assessment of numerical accuracy through the numerical standard errors. Relative numerical efficiency for the coefficient posterior means ranges from 0.23 to 0.29: overall, the client obtains the same numerical accuracy he would have achieved executing the simulator directly with about 2500 iterations. This would have required \( 2.5 \text{ sec} \) and the use of the \texttt{minst} and \texttt{postsim} commands, whereas the reweighting took about \( 3.5 \text{ sec} \) and the simpler \texttt{expect1} command. Reweighting of the simulator output holds little advantage in this example due to the simplicity of simulating the regression model.

A similar exercise was conducted for the probit example introduced in Section 3.8. To mimic what an investigator might do, a posterior simulation structure for the posterior distribution, with a prior distribution, in which all standard deviations were 10 times larger than those indicated in Table 2, was created. This simulator output was then reweighted to reflect the prior distribution in Table 2. This attempt failed completely—two draws received more than 99.99% of the total weight. To appreciate the reason for this failure recall that the prior distribution used in Section 3.8 is highly informative relative to the sample. This is evident from inspection of Table 2.

To complete this example, the investigator was recast in the role of a client—i.e., the client now has the priors indicated in Table 2 except that the standard deviations are 10 times larger than shown there. Imagine an investigator who uses a prior distribution with standard deviations 10 times larger yet—i.e., the standard deviations are 100 times those shown for the prior distribution in Table 2. This client’s reweighting of this investigator’s simulator output yields the results displayed in the right half of Table 7. The chi-squared statistic (again
<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Posterior (tightest prior, direct Gibbs)</th>
<th>Posterior (reweighting of MCMC from more diffuse prior)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Stand. dev.</td>
</tr>
<tr>
<td>Intercept</td>
<td>7.7283</td>
<td>0.2144</td>
</tr>
<tr>
<td>Driveway</td>
<td>0.10908</td>
<td>0.02465</td>
</tr>
<tr>
<td>Rec room</td>
<td>0.069476</td>
<td>0.02307</td>
</tr>
<tr>
<td>Fin basement</td>
<td>0.10304</td>
<td>0.01950</td>
</tr>
<tr>
<td>Gas hot water</td>
<td>0.14254</td>
<td>0.03316</td>
</tr>
<tr>
<td>Central air</td>
<td>0.15408</td>
<td>0.01961</td>
</tr>
<tr>
<td>#Garage stalls</td>
<td>0.051947</td>
<td>0.01119</td>
</tr>
<tr>
<td>Good nbhd</td>
<td>0.12561</td>
<td>0.02043</td>
</tr>
<tr>
<td>log(Lot size)</td>
<td>0.30469</td>
<td>0.02625</td>
</tr>
<tr>
<td>#Bedrooms</td>
<td>0.040706</td>
<td>0.013615</td>
</tr>
<tr>
<td>#Full baths</td>
<td>0.15519</td>
<td>0.018767</td>
</tr>
<tr>
<td>#Stories</td>
<td>0.083422</td>
<td>0.012002</td>
</tr>
<tr>
<td>Coefficient</td>
<td>Posterior (Gibbs, direct MCMC)</td>
<td>Posterior (reweighting of MCMC from more diffuse prior)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Stand. dev.</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.545</td>
<td>0.4618</td>
</tr>
<tr>
<td>Black</td>
<td>0.1072</td>
<td>0.1048</td>
</tr>
<tr>
<td>Age-Single</td>
<td>-0.05702</td>
<td>0.01208</td>
</tr>
<tr>
<td>Age-Married</td>
<td>-0.08414</td>
<td>0.02247</td>
</tr>
<tr>
<td>Education</td>
<td>0.07848</td>
<td>0.02247</td>
</tr>
<tr>
<td>Married</td>
<td>0.6779</td>
<td>0.4744</td>
</tr>
<tr>
<td>Kids</td>
<td>-0.5035</td>
<td>0.1746</td>
</tr>
<tr>
<td>#Kids</td>
<td>-0.05377</td>
<td>0.05461</td>
</tr>
<tr>
<td>Spouse$</td>
<td>-1.702 \times 10^{-6}</td>
<td>3.436 \times 10^{-6}</td>
</tr>
<tr>
<td>Family$</td>
<td>1.043 \times 10^{-6}</td>
<td>6.397 \times 10^{-6}</td>
</tr>
<tr>
<td>AFDC</td>
<td>-5.342 \times 10^{-4}</td>
<td>3.782 \times 10^{-4}</td>
</tr>
<tr>
<td>Food$</td>
<td>-0.002225</td>
<td>6.946 \times 10^{-4}</td>
</tr>
<tr>
<td>WorkExp</td>
<td>1.372 \times 10^{-4}</td>
<td>9.392 \times 10^{-6}</td>
</tr>
</tbody>
</table>
from \texttt{expectN}) for comparison of the posterior means again indicates no problem with numerical standard errors (NSEs). The relative numerical efficiencies (RNEs) show that the investigator’s simulator output with 10,000 records provides about the same information the client would have obtained with 1000 records directly from simulator output for his posterior. This would have required the \texttt{minst} and \texttt{postsim} commands and about 40 sec of execution time, whereas the reweighting required less than 4 sec and the simpler \texttt{expect1} command. (Notice that the posterior means, in Table 7, are much closer to the maximum likelihood statistics in Table 2 than are the posterior means in the same table, which correspond to the previous more informative prior.)

These examples underscore both the potential efficiency of Bayesian communication through posterior reweighting, and its limitations. The efficiency comes about because reweighting software is simple and generic, whereas posterior simulators are model specific and impose greater computational demands. (This advantage increases dramatically in more complex models.) The limitations arise from the need for some similarity of the investigator’s and client’s posterior distributions. As argued and illustrated here, the investigator should use a prior distribution that is uninformative relative to the sample. In this situation a client’s reweighting will be successful for priors that are moderately, but not greatly, informative relative to the sample.

\section*{ACKNOWLEDGMENT}

This chapter is a revision of the article “Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication” that appeared, together with comments and reply, as issue no. 1 in volume 18 of \textit{Econometric Reviews}, early in 1999. The illustrations in this chapter use a more recent version of BACC than did the earlier article, and the article treats aspects of model development that are not included in this chapter. Financial support from NSF grants SBR-9600040, SBR-9731037, and SBR-0214303 is gratefully acknowledged.

\section*{REFERENCES}


1 INTRODUCTION

Zellner's idea of combining several equations into one model to improve estimation efficiency (Zellner, 1962) ranks as one of the most successful and lasting innovations in the history of econometrics. The resulting seemingly unrelated regressions (SUR) model has generated a wealth of both theoretical and empirical contributions. Reviews of work on or involving the SUR model can be found in Srivastava and Dwivedi (1979), Judge et al. (1985), Srivastava and Giles (1987), and Fiebig (2001). It was also Zellner (1971) who popularized Bayesian inference in econometrics generally and described the SUR model within the context of Bayesian inference. However, at that time, convenient methods for deriving or estimating marginal posterior density functions and moments for individual SUR coefficients were not generally available. Subsequently, analytical results were derived for some special cases (Drèze and Morales, 1976; Richard and Tompa, 1980; Richard and Steel, 1988; Steel 1992) and importance sampling was suggested as a means for estimating marginal posterior density functions and their moments (Kloek and van Dijk, 1978). More recently, the application of Markov-Chain Monte Carlo (MCMC) methodology to Bayesian inference has made available a new range of
numerical methods that make Bayesian estimation of the SUR model more convenient and accessible. The literature of MCMC is extensive; for a general appreciation of its scope and purpose, see Tierney (1994), Albert and Chib (1996), Chen et al. (2000), Chib and Greenberg (1996), Gilks et al. (1996), Tanner (1996), and the chapter by Geweke et al. (this volume). For application of MCMC to the SUR model, see, e.g., Percy (1992, 1996), Chib and Greenberg (1995), Griffiths and Chotikapanich (1997), and Griffiths et al. (2000).

The objective of this chapter is to provide a practical guide to computer-aided Bayesian inference for a variety of problems that arise in applications of the SUR model. We describe examples of problems, models, and algorithms that have been placed within a general framework in the chapter by Geweke et al. (this volume); our chapter can be viewed as complementary to that one. The model is described in Section 2; the joint, conditional, and marginal posterior density functions that result from a noninformative prior are derived. In Section 3 we describe how to use sample draws of parameters from their posterior densities to estimate posterior quantities of interest; two Gibbs sampling algorithms and a Metropolis–Hastings algorithm are given. Modifications necessary for nonlinear equations, equality restrictions and inequality restrictions are presented in Sections 4, 5, and 6, respectively. Three applications are described in Section 7. Section 8 contains methodology for forecasting. Some extensions are briefly mentioned in Section 9 and a few concluding remarks are given in Section 10.

2 MODEL SPECIFICATION AND POSTERIORS FROM A NONINFORMATIVE PRIOR

Consider $M$ equations written as

$$y_i = X_i \beta_i + e_i \quad i = 1, 2, \ldots, M$$

where $y_i$ is a $T$-dimensional vector of observations on a dependent variable, $X_i$ is a $(T \times K_i)$ matrix of observations on $K_i$ nonstochastic explanatory variables, possibly including a constant term, $\beta_i$ is a $K_i$-dimensional vector of unknown coefficients that we wish to estimate, and $e_i$ is a $T$-dimensional unobserved random vector. The $M$ equations can be combined into one large model written as

$$
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_M
\end{bmatrix} = 
\begin{bmatrix}
X_1 & X_2 & \cdots & X_M
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_M
\end{bmatrix} + 
\begin{bmatrix}
e_1 \\
e_2 \\
\vdots \\
e_M
\end{bmatrix}
$$

(2)
which we then write compactly as

$$ y = X\beta + e $$  

(3)

where \( y \) is of dimension \( (TM \times 1) \), \( X \) is of dimension \( (TM \times K) \), with \( K = \sum_{i=1}^{M} K_i \), \( \beta \) is \( (K \times 1) \) and \( e \) is \( (TM \times 1) \). We assume that the distribution for \( e \) is given by

$$ e \sim N(0, \Sigma \otimes I_T) $$  

(4)

Thus, the errors in each equation are homoskedastic and not autocorrelated. There is, however, contemporaneous correlation between corresponding errors in different equations. The variance in the error of the \( i \)th equation we denote by \( \sigma_{ii} \), the \( i \)th diagonal element of \( \Sigma \). The covariance between two corresponding errors in different equations (say \( i \) and \( j \)), we write as \( \sigma_{ij} \), an off-diagonal element of \( \Sigma \).

Using \( f(\cdot) \) as generic notation for a probability density function (p.d.f.), the likelihood function for \( \beta \) and \( \Sigma \) can be written as

$$ f(y|\beta, \Sigma) = (2\pi)^{-MT/2}|\Sigma|^{-T/2} \exp\left\{-\frac{1}{2}(y - X\beta)'(\Sigma^{-1} \otimes I_T)(y - X\beta)\right\} $$

(5)

This p.d.f. can also written as

$$ f(y|\beta, \Sigma) = (2\pi)^{-MT/2}|\Sigma|^{-T/2} \exp\left\{-\frac{1}{2}\text{tr}(A\Sigma^{-1})\right\} $$

(6)

where \( A \) is an \((M \times M)\) matrix with \((i,j)\)th element given by

$$ [A]_{ij} = (y_i - X_i\beta_i)(y_j - X_j\beta_j) $$

(7)

Note that \( A \) can also be written as

$$ A = (Y - X^*B)'(Y - X^*B) $$

(8)

where \( Y \) is the \((T \times M)\) matrix \( Y = (y_1, y_2, \ldots, y_M) \), \( X^* \) is the \((T \times K)\) matrix \( X^* = (X_1, X_2, \ldots, X_M) \), and \( B \) is the \((K \times M)\) matrix:

$$ B = \begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_M
\end{bmatrix} $$

(9)

Result (9) on page 42 of Lütkepohl (1996) can be used to establish the equivalence of Eqs. (5) and (6). Specifically,

$$ \text{tr}[(Y - X^*B)'(Y - X^*B)\Sigma^{-1}] = [\text{vec}(Y - X^*B)]' [\Sigma^{-1} \otimes I_T] \text{vec}(Y - X^*B) $$

$$ = (y - X\beta)'(\Sigma^{-1} \otimes I_T)(y - X\beta) $$

(10)
Two prior p.d.f.s. will be considered in this chapter; they are the conventional noninformative prior (see, e.g., Zellner 1971, ch. 8):

\[ f(\beta, \Sigma) = f(\beta) f(\Sigma) \propto |\Sigma|^{-(M+1)/2} \]  

(11)

and another prior that imposes inequality restrictions on \( \beta \), but is otherwise noninformative. The inequality prior and its consequences will be considered later in the chapter. The noninformative prior in Eq. (11) is chosen to provide objectivity in reporting, not because we believe total ignorance is prevalent. Geweke et al. (Chap. 8 in this volume) discuss how to modify results to accommodate the prior of a specific client.

2.1 Joint Posterior p.d.f. for \((\beta, \Sigma)\)

Applying Bayes’ theorem to the prior p.d.f. in Eq. (11) and the likelihood function in Eqs. (5) and (6) yields the joint posterior p.d.f. for \( \beta \) and \( \Sigma \):

\[ f(\beta, \Sigma | y) \propto f(y | \beta, \Sigma) f(\beta, \Sigma) \]

\[ \propto |\Sigma|^{-(T+M+1)/2} \exp\left\{ -\frac{1}{2} (y - X\beta)'(\Sigma^{-1} \otimes I_T)(y - X\beta) \right\} \]

\[ = |\Sigma|^{-(T+M+1)/2} \exp\left\{ -\frac{1}{2} \text{tr}(A \Sigma^{-1}) \right\} \]  

(12)

In the remainder of this section we describe a number of marginal and conditional posterior p.d.f.’s that are derived from Eq. (12). These p.d.f.’s will prove useful in later sections when we discuss methods for estimating quantities of interest. We will assume that interest centers on individual coefficients, say the \( k \)th coefficient in the \( i \)th equation \( \beta_{ik} \), and, more generally, on some functions of the coefficients, say \( g(\beta) \). Forecasting future values \( y^* \) will also be considered. The relevant p.d.f.’s that express our uncertain postsample knowledge about these quantities are the marginal p.d.f.’s \( f(\beta_{ik} | y) \), \( f(g(\beta) | y) \), and \( f(y^* | y) \), respectively. Typically, we report results by graphing these p.d.f.’s, and tabulating their means, standard deviations, and probabilities for regions of interest. Describing the tools for doing so is the major focus of this chapter.

2.2 Conditional Posterior p.d.f. for \((\beta | \Sigma)\)

The term in the exponent of Eq. (12) can be written as

\[ (y - X\beta)'(\Sigma^{-1} \otimes I_T)(y - X\beta) = (y - X\hat{\beta})'(\Sigma^{-1} \otimes I_T)(y - X\hat{\beta}) \]

\[ + (\beta - \hat{\beta})'X'(\Sigma^{-1} \otimes I_T)X(\beta - \hat{\beta}) \]  

(13)
where $\hat{\beta} = [X'(\Sigma^{-1} \otimes I_T)X]^{-1}X'(\Sigma^{-1} \otimes I_T)y$. It follows that the conditional posterior p.d.f. for $\beta$ given $\Sigma$ is the multivariate normal p.d.f.:

$$f(\beta | \Sigma, y) \propto \exp\left\{-\frac{1}{2}(\beta - \hat{\beta})'X'(\Sigma^{-1} \otimes I_T)X(\beta - \hat{\beta})\right\}$$

with posterior mean equal to the generalized least squares estimator:

$$E(\beta | y, \Sigma) = \hat{\beta} = [X'(\Sigma^{-1} \otimes I_T)X]^{-1}X'(\Sigma^{-1} \otimes I_T)y$$

and posterior covariance matrix equal to

$$V(\beta | y, \Sigma) = [X'(\Sigma^{-1} \otimes I_T)X]^{-1}$$

The last two expressions are familiar ones in sampling theory inference for the SUR model. They show that the traditional SUR estimator, written as

$$\hat{\beta} = [X'(\Sigma^{-1} \otimes I_T)X]^{-1}X'(\Sigma^{-1} \otimes I_T)y$$

where $\hat{\Sigma}$ is a two-step estimator or a maximum likelihood estimator, can be viewed as the mean of the conditional posterior p.d.f. for $\beta$ given $\Sigma$. The traditional covariance matrix estimator $[X'(\hat{\Sigma}^{-1} \otimes I_T)X]^{-1}$ can be viewed as the conditional covariance matrix from the same p.d.f. Since this p.d.f. does not take into account uncertainty from not knowing $\Sigma$ (the fact that $\hat{\Sigma}$ is an estimate is not recognized), it overstates the reliability of our information about $\beta$. This dilemma was noted by Fiebig and Kim (2000) in the context of an increasing number of equations.

### 2.3 Marginal Posterior p.d.f. for $\beta$

The more appropriate representation of our uncertainty about $\beta$ is the marginal posterior p.d.f. $f(\beta | y)$. It can be shown that this p.d.f. is given by

$$f(\beta | y) = \int f(\beta, \Sigma | y) d\Sigma \propto |A|^{-T/2}$$

The integral in Eq. (18) is performed by using properties of the inverted Wishart distribution (see, e.g., Zellner 1971, p. 395). For $f(\beta | y) \propto |A|^{-T/2}$ to be proper, we require $T \geq M + \text{rank}(X^T)$ (Griffiths et al., 2002). Also, this p.d.f. is not of a standard recognizable form. Except for special cases, analytical expressions for its normalizing constant and moments are not available. Estimating these moments, and marginal p.d.f.’s for individual coefficients $\beta_{ik}$, is considered in Section 3; first, we describe some more p.d.f.’s that will prove to be useful.
2.4 Conditional Posterior p.d.f. for \((\beta_1 \mid \beta_2, \ldots, \beta_M)\)

It is possible to show that the posterior p.d.f. for the coefficient vector from one equation, conditional on those from other equations, is a multivariate \(t\)-distribution.

To derive this result, we will consider the posterior p.d.f. for \(\beta_1\), conditional on \((\beta_2, \beta_3, \ldots, \beta_M)\). We write a partition of \((Y - X^*B)\) into its first and remaining \((M - 1)\) columns as

\[ Y - X^*B = (y_1 - X_1\beta_1 \quad E_{(1)}) \]

The corresponding partition of \(A\) is

\[ A = \begin{bmatrix} (y_1 - X_1\beta_1)'(y_1 - X_1\beta_1) & (y_1 - X_1\beta_1)'E_{(1)} \\ E_{(1)}'(y_1 - X_1\beta_1) & E_{(1)}'E_{(1)} \end{bmatrix} \]

Using a result on the determinant of a partitioned matrix, we have

\[ |A| = |E_{(1)}E_{(1)}'|((y_1 - X_1\beta_1)'(y_1 - X_1\beta_1) \]

\[- (y_1 - X_1\beta_1)'E_{(1)}(E_{(1)}'E_{(1)})^{-1}E_{(1)}'(y_1 - X_1\beta_1)) \]

Defining \(Q_{(1)} = I_T - E_{(1)}E_{(1)}'E_{(1)}^{-1}E_{(1)}'\) and \(\tilde{\beta}_1 = (X_1'Q_{(1)}X_1)^{-1}X_1'Q_{(1)}y_1\), the second term in the above equation can be written as

\[ (y_1 - X_1\beta_1)'Q_{(1)}(y_1 - X_1\beta_1) = (y_1 - X_1\tilde{\beta}_1)'Q_{(1)}(y_1 - X_1\tilde{\beta}_1) \]

\[ + (\beta_1 - \tilde{\beta}_1)'X_1'Q_{(1)}X_1(\beta_1 - \tilde{\beta}_1) \]

Collecting all these results, substituting into Eq. (18), and letting \(|E_{(1)}'E_{(1)}|\) be absorbed into the proportionality constant, we can write

\[ f(\beta_1 \mid y, \beta_2, \beta_3, \ldots, \beta_M) \propto \left[ v_1 + \frac{(\beta_1 - \tilde{\beta}_1)'X_1'Q_{(1)}X_1(\beta_1 - \tilde{\beta}_1)}{s_i^2} \right]^{-(K_1 + v_1)/2} \]

(19)

where \(v_1 = T - K_1\) and \(s_i^2 = (y_1 - X_1\tilde{\beta}_1)'Q_{(1)}(y_1 - X_1\tilde{\beta}_1)/v_1\). Equation (19) is in the form of a multivariate \(t\)-distribution with degrees of freedom \(v_1\), mean \(\tilde{\beta}_1\), and covariance matrix \((v_1/(v_1 - 2))s_i^2(X_1'Q_{(1)}X_1)^{-1}\). See, e.g., Zellner (1971, p. 383). The conditional posterior p.d.f.'s for other \(\beta_j\) are similarly defined.
2.5 Conditional Posterior p.d.f. for \((\Sigma | \beta)\)

Viewing the joint posterior p.d.f. in Eq. (12) as a function of only \(\Sigma\) yields the conditional posterior p.d.f. for \(\Sigma\) given \(\beta\). It is the inverted Wishart p.d.f. (see, e.g., Zellner, 1971, p. 395):

\[
f(\Sigma | \beta, y) \propto |\Sigma|^{-(T+M+1)/2} \exp\left(-\frac{1}{2} \text{tr}(A\Sigma^{-1})\right)
\]

It has \(T\) degrees of freedom, and parameter matrix \(A\).

2.6 Marginal Posterior pdf for \(\Sigma\)

The marginal p.d.f. for \(\Sigma\), obtained by using the result in Eq. (13), and then using properties of the multivariate normal p.d.f. to integrate out \(\beta\), is given by

\[
f(\Sigma | y) = \int f(\beta, \Sigma | y) \, d\beta
\]

\[
= |X'(\Sigma^{-1} \otimes I_T)X|^{-1/2} |\Sigma|^{-(T+M+1)/2} \times \exp\left\{-\frac{1}{2} (y - X\hat{\beta})'(\Sigma^{-1} \otimes I_T)(y - X\hat{\beta})\right\}
\]

\[
= |X'(\Sigma^{-1} \otimes I_T)X|^{-1/2} |\Sigma|^{-(T+M+1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\hat{A}\Sigma^{-1})\right\}
\]

where \(\hat{A}\) is an \((M \times M)\) matrix with the \((i,j)\)th element given by

\[
[A]_{ij} = (y_i - X_i\hat{\beta}_i)(y_j - X_j\hat{\beta}_j)
\]

The posterior p.d.f. in Eq. (21) is not an analytically tractable one whose moments are known. However, as we will see, we can draw observations from it using the Gibb’s sampler.

3 ESTIMATING POSTERIOR QUANTITIES

Given the intractability of the posterior p.d.f. \(f(\beta | y) \propto A^{-T/2}\), methods for estimating marginal posterior p.d.f.’s for individual coefficients \(\beta_{ik}\), their moments, and probabilities of interest, are required. Suppose that we have draws \(\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(N)}\) taken from \(f(\beta | y)\) and, possibly, draws \(\Sigma^{(1)}, \Sigma^{(2)}, \ldots, \Sigma^{(N)}\) taken from \(f(\Sigma | y)\). We will describe a number of ways one can proceed to estimate the desired quantities; then, we discuss how the required posterior draws can be obtained.

3.1 Estimating Posterior p.d.f.s

A simple way to estimate the marginal posterior p.d.f. of \(\beta_{ik}\), say, is to construct a histogram of draws of that parameter. Joining the midpoints
of the histogram classes provides a continuous representation of the p.d.f., but typically, it will be jagged one unless some kind of smoothing procedure is employed. Alternatively, one can obtain a smooth p.d.f., and a more efficient estimate, by averaging conditional posterior p.d.f's for the quantity of interest. In this case, for conditional posterior p.d.f’s one can use the \( t \)-distributions defined by Eq. (19), or, if draws on both \( \beta \) and \( \Sigma \) are available, the normal distributions defined by Eq. (14).

Considering the \( t \)-distribution first, an estimate of \( f(\beta_{ik} | y) \) is given by

\[
\hat{f}(\beta_{ik} | y) = \frac{1}{N} \sum_{\ell=1}^{N} f(\beta_{ik} | y, \beta_{i(o)}^{(\ell)}, \ldots, \beta_{i(o)}^{(M)}) \\
= c \frac{1}{N} \sum_{\ell=1}^{N} \frac{1}{\sqrt{s_i^{(\ell)}}} \frac{1}{q_{i(k)k}^{(\ell)}} \left[ v_i + \frac{(\beta_{ik} - \hat{\beta}_{ik}^{(\ell)})^2}{s_i^{(\ell)} q_{i(k)k}^{(\ell)}} \right]^{-(v_i+1)/2}
\]  

(22)

The univariate \( t \)-distribution that is being averaged in Eq. (22) is the conditional p.d.f. for a single coefficient from \( \beta_{i(o)} \), obtained from the multivariate \( t \)-distribution in Eq. (19), after generalizing from \( \beta_{i(o)} \) to \( \beta_{i(k)} \).

The previously undefined terms in Eq. (22) are the constant:

\[
c = \frac{\Gamma[(v + 1)/2] v^{v/2}}{\Gamma(v/2) \sqrt{\pi}}
\]

where \( \Gamma(\cdot) \) is the gamma function, the conditional posterior mean \( \hat{\beta}_{ik} \), which is the \( k \)-element in \( \hat{\beta} \), and \( q_{i(k)k} \), which is the \( k \)th diagonal element of \( (X_i' Q_i X_i)^{-1} \). To plot the p.d.f. in Eq. (22), we choose a grid of values for \( \beta_{ik} \) (50–100 is usually adequate), and for each value of \( \beta_{ik} \), we compute the average in Eq. (22). These averages are plotted against the \( \beta_{ik} \).

Alternatively, the conditional normal distributions in Eq. (14) can be averaged over \( \Sigma \). In this case an estimate of the marginal posterior p.d.f. for \( \beta_{ik} \) is given by

\[
\hat{f}(\beta_{ik} | y) = \frac{1}{N} \sum_{\ell=1}^{N} f(\beta_{ik} | y, \Sigma^{(\ell)}) \\
= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{h_{i(k)k}^{(\ell)}}} \exp \left\{ -\frac{1}{2h_{i(k)k}^{(\ell)}} (\beta_{ik} - \hat{\beta}_{ik}^{(\ell)})^2 \right\}
\]  

(23)

where \( \hat{\beta}_{ik} \) is the \( k \)-element in the \( i \)th vector component of \( \hat{\beta} \) [see Eq. (15)] and \( h_{i(k)k} \) is the \( k \)-diagonal element in the \( i \)th diagonal block of \( [X'(\Sigma^{-1} \otimes I_T)X]^{-1} \) [see Eq. (16)]. As in Eq. (22), the average in Eq. (23) is computed for, and plotted against, a grid of values for \( \beta_{ik} \).
3.2 Estimating Posterior Means and Standard Deviations

Corresponding to the three ways given for estimating posterior p.d.f.’s, there are three ways of estimating their posterior means and variances. The first way is to use the sample mean and covariance matrix of the draws. That is,

\[ \hat{E}(\beta | y) = \frac{1}{N} \sum_{\ell=1}^{N} \beta^{(t)} = \bar{\beta} \]  

(24)

and

\[ \hat{V}(\beta | y) = \frac{1}{N - 1} \sum_{\ell=1}^{N} (\beta^{(t)} - \bar{\beta})(\beta^{(t)} - \bar{\beta})' \]  

(25)

The second and third approaches use the results: (1) an unconditional mean is equal to the mean of the conditional means, and (2) the unconditional variance is equal to the mean of the conditional variances plus the variance of the conditional means. Applying these two results to the conditional posterior p.d.f. in Eq. (19) yields:

\[ \hat{E}(\beta_i | y) = \frac{1}{N} \sum_{\ell=1}^{N} \beta_i^{(t)} = \frac{1}{N} \sum_{\ell=1}^{N} (X_i'Q_{(i)}^oX_i)^{-1}X_i'Q_{(i)}^o y_i = \bar{\beta}_i \]  

(26)

\[ \hat{V}(\beta_i | y) = \left( \frac{V_i}{V_i - 2} \right) \frac{1}{N} \sum_{\ell=1}^{N} s_i^{2(t)} (X_i'Q_{(i)}^oX_i)^{-1} \right. 

+ \frac{1}{N - 1} \sum_{\ell=1}^{N} (\beta_i^{(t)} - \bar{\beta}_i)(\beta_i^{(t)} - \bar{\beta}_i)' \]  

(27)

Applying the two results to the normal conditional posterior p.d.f.’s in Eq. (14) yields:

\[ \hat{E}(\beta | y) = \frac{1}{N} \sum_{\ell=1}^{N} \beta^{(t)} = \frac{1}{N} \sum_{\ell=1}^{N} [X'(\Sigma^{-1(t)} \otimes I_T)X]^{-1}X'(\Sigma^{-1(t)} \otimes I_T)y = \bar{\beta} \]  

(28)

and

\[ \hat{V}(\beta | y) = \frac{1}{N} \sum_{\ell=1}^{N} [X'(\Sigma^{-1(t)} \otimes I_T)X]^{-1} \right. 

+ \frac{1}{N - 1} \sum_{\ell=1}^{N} (\beta^{(t)} - \bar{\beta})(\beta^{(t)} - \bar{\beta})' \]  

(29)
Clearly, using the sample means and standard deviations from Eqs. (24) and (25) is much easier than using the conditional quantities in Eqs. (26)–(29). However, averaging conditional moments generally leads to more efficient estimates.

### 3.3 Estimating Probabilities

Often, we are interested in reporting the probability that $\beta_{ik}$ lies with a particular interval or finding an interval with a prespecified probability content. In sampling theory, inference intervals with 95% probability content are popular. An estimate of the probability that $\beta_{ik}$ lies in a particular interval is given by the proportion of draws that lie within that interval. Alternatively, one can find conditional probabilities and average them, along the lines that the conditional means are averaged in Eq. (26) and (28). Using the conditional normal distribution as an example, we can estimate the probability that $\beta_{ik}$ lies in the interval $(a, b)$ as

$$
\hat{P}(a < \beta_{ik} < b) = \frac{1}{N} \sum_{\ell=1}^{N} P(a < \beta_{ik} | \Sigma^{(\ell)}) < b)
$$

Order statistics can be used to obtain an interval with a prespecified probability content. For example, for a 95% probability interval for $\beta_{ik}$, we can take the 0.025 and 0.975 empirical quantiles of the draws of the $\beta_{ik}$.

### 3.4 Functions of $\beta$

How do we proceed if we are interested in some functions of $\beta$, say $g(\beta)$? Examples of such functions considered later in this chapter are monotonicity and curvature conditions in a cost system, and the relative magnitudes of equivalence scales in household expenditure functions. Examples outside the context of SUR models are the evaluation of consumer surplus (Griffiths, 1999) and the stationary region in a time-series model (Geweke, 1988).

If it is possible to derive, analytically, conditional distributions of the form $f(g(\beta_{1}) | y, \beta_{2}, \ldots, \beta_{M})$ or $f(g(\beta) | y, \Sigma)$, then one can work with these conditional distributions along the lines described above. However, the ability to proceed analytically is rare, given that $g(\beta)$ is frequently nonlinear and of lower dimension than $\beta$. Instead, we can compute values $g(\beta^{(\ell)})$, $\ell = 1, 2, \ldots, N$, from the draws of $\beta$. These values can be placed in a histogram to estimate the p.d.f. of $g(\beta)$. Their sample mean and variance can be used to estimate the corresponding posterior mean and variance. Probabilities can be estimated using the proportion of values in a given
region and order statistics can be used to find an interval with a given probability content.

3.5 Gibbs Sampling with $\beta$ and $\Sigma$

We now turn to the question of how to obtain draws $\beta$ and $\Sigma$ from their respective marginal posterior p.d.f.’s. One possible way is to use an MCMC algorithm known as Gibbs sampling. In this procedure draws are made iteratively from the conditional posterior p.d.f.’s. Specifically, given a particular starting value for $\Sigma$, say $\Sigma^{(0)}$, the $\ell$th draw from the Gibbs sampler $(\beta^{(\ell)}, \Sigma^{(\ell)})$ is obtained using following two steps:

1. Draw $\beta^{(\ell)}$ from $f(\beta | \Sigma^{(\ell-1)}, y)$
2. Draw $\Sigma^{(\ell)}$ from $f(\Sigma | \beta^{(\ell)}, y)$

Making these draws is straightforward, given that the two conditional posterior p.d.f.’s are normal and inverted Wishart, respectively. (See Appendix for details.) MCMC theory suggests that, after a sufficiently large number of draws, the Markov Chain created by the draws will converge. After convergence, the subsequent draws can be viewed as draws from the marginal posterior p.d.f.’s $f(\beta | y)$ and $f(\Sigma | y)$. It is these draws that can be used to present results in the desired fashion. Draws taken prior to the point at which convergence is assumed to have taken place are sometimes called the “burn in”; they are discarded. A large number of diagnostics have been suggested for assessing whether convergence has taken place. See, e.g. Cowles and Carlin (1996). Assessing whether convergence has taken place is similar to assessing whether a time series is stationary. Thus, visual inspection of a graph of the sequence of draws, testing whether the mean and variance are the same at the beginning of the chain as at the end of the chain, and testing whether two or more separately run chains have the same mean and variance, are ways of checking for convergence.

Since we are using a sample of draws of $\beta$ and $\Sigma$ to estimate posterior means and standard deviations, and other relevant population quantities, the accuracy of the estimates is a concern. Estimation accuracy is assessed using numerical standard errors. Methods for computing such standard errors are described in Chapter 8 of this volume by Geweke et al. Because the draws produced by MCMC algorithms are correlated, time-series methods are used to compute the standard errors; also larger samples are required to achieve a given level of accuracy relative to a situation involving independent draws.
Although the above remarks on convergence and numerical standard errors were made in the context of the Gibbs sampler for $\beta$ and $\Sigma$, they also apply to other MCMC algorithms including the Gibbs sampler for $\beta$ and the Metropolis–Hastings algorithm described below.

### 3.6 Gibbs Sampling with $\beta$

If the number of equations is large, making $\Sigma$ of high dimension, then it may be preferable to use a Gibbs sampler based on the conditional posterior p.d.f.’s for the $\beta_i$ from each equation. Note, however, that this alternative is not feasible if cross-equation restrictions on the $\beta_i$, as discussed in Sections 5 and 7, are present.

To proceed with this Gibbs sampler, we begin with starting values for all coefficients except the first, say $\beta_1^{(0)}, \beta_2^{(0)}, \ldots, \beta_M^{(0)}$, and then sample iteratively using the following steps for the $t$th draw:

1. Draw $\beta_1^{(t)}$ from $f(\beta_1 | \beta_2^{(t-1)}, \ldots, \beta_M^{(t-1)})$
2. Draw $\beta_2^{(t)}$ from $f(\beta_2 | \beta_1^{(t)}, \beta_3^{(t-1)}, \ldots, \beta_M^{(t-1)})$
   
   \[ \vdots \]

   \[ i. \text{ Draw } \beta_i^{(t)} \text{ from } f(\beta_i | \beta_1^{(t)}, \ldots, \beta_i^{(t-1)}, \beta_{i+1}^{(t-1)}, \ldots, \beta_M^{(t-1)}) \]
   
   \[ \vdots \]

   \[ M. \text{ Draw } \beta_M^{(t)} \text{ from } f(\beta_M | \beta_1^{(t)}, \ldots, \beta_M^{(t-1)}) \]

The conditional posterior p.d.f.’s are multivariate $t$-distributions from which we can readily draw values (see Appendix). Ordinary least squares estimates are adequate for starting values.

### 3.7 Metropolis–Hastings Algorithm

An alternative to Gibbs sampling is a Metropolis–Hastings algorithm that draws observations from the marginal posterior p.d.f. $f(\beta | y)$. As we shall see, this algorithm is particularly useful for an inequality-restricted prior, or if the equations are nonlinear. The algorithm we describe is a random-walk algorithm; it is just one of many possibilities. For others see, e.g., Chen et al. (2000).

The Metropolis–Hastings algorithm generates a candidate value $\beta^*$ that is accepted or rejected as a draw from the posterior p.d.f. $f(\beta | y)$. When it is rejected, the previously accepted draw is repeated as a draw. Thus, rules are needed for generating the candidate value $\beta^*$ and for accepting it. Let $V$ be the covariance matrix for the distribution used to generate a candidate value. The maximum likelihood covariance matrix is usually suitable. For the linear SUR model this matrix is $[X'(V^{-1} \otimes I_p)X]^{-1}$. Choose a feasible starting value $\beta^{(0)}$. The following
steps can be used to draw the \((\ell + 1)\)th observation in a random-walk
Metropolis–Hastings chain:

1. Draw a candidate value \(\beta^*\) from an \(N(\beta^{(\ell)}, cV)\) distribution where
   \(c\) is a scalar set such that \(\beta^*\) is accepted approximately 40–50% of the time.
2. Compute the ratio of the posterior p.d.f. evaluated at the
   candidate draw to the posterior p.d.f. evaluated at the previously
   accepted draw:
   \[
   r = \frac{f(\beta^* | y)}{f(\beta^{(\ell)} | y)}
   \]
   Note that this ratio can be computed without knowledge of the
   normalizing constant for \(f(\beta | y)\). Also, if any of the elements of \(\beta^*\)
   fall outside a feasible parameter region defined by an inequality-
   restricted prior (see Section 5), then \(f(\beta^* | y) = 0\). When \(r > 1\), \(\beta^*\) is
   a more likely value then \(\beta^{(\ell)}\) in the sense that it is closer to the mode
   of the distribution. When \(r < 1\), \(\beta^*\) is further into the tails of the
   distribution. If \(r > 1\), \(\beta^*\) is accepted; if \(r < 1\), \(\beta^*\) is accepted with
   probability \(r\). Thus, more draws occur in regions of high
   probability and fewer draws occur in regions of low probability.
   Details of the acceptance–rejection procedure follow in step 3.
3. Draw a value \(u\) for a uniform random variable on the interval
   \((0, 1)\):
   \[
   \begin{align*}
   &\text{If } u \leq r, \text{ set } \beta^{(\ell+1)} = \beta^* \\
   &\text{If } u > r, \text{ set } \beta^{(\ell+1)} = \beta^{(\ell)} \\
   &\text{Return to step 1 with } \ell \text{ set to } \ell + 1.
   \end{align*}
   \]
   Let \(q(\beta^* | \beta^{(\ell)})\) be the distribution used to generate the candidate value \(\beta^*\) in
   step 1. In our case it is a normal distribution. In more general Metropolis–
   Hastings algorithms, where our choice of distribution is not necessarily
   utilized, \(r\) is defined as
   \[
   r = \frac{f(\beta^* | y) \times q(\beta^{(\ell)} | \beta^*)}{f(\beta^{(\ell)} | y) \times q(\beta^* | \beta^{(\ell)})}
   \]
   In our case \(q(\beta^{(\ell)} | \beta^*) = q(\beta^* | \beta^{(\ell)})\). Various alternatives for \(q()\) have been
   suggested in the literature.

4 NONLINEAR SUR

Many economic models are intrinsically nonlinear, or a nonlinear model
may result from substituting nonlinear restrictions on \(\beta\) into a linear
model. The Gibbs sampling algorithms that we described are no longer applicable for a nonlinear SUR model. However, we can still proceed with the Metropolis–Hastings algorithm.

Suppose that the nonlinear SUR model is given by

\[ y_i = h_i(X, \beta) + e_i \quad i = 1, 2, \ldots, M \]  

where the \( h_i \) are the nonlinear functions. The dimensions of \( y_i, h_i, \) and \( e_i \) are \((T \times 1)\). In this context \( X \) represents a set of explanatory variables and \( \beta \) is the vector of all unknown coefficients. The omission of an \( i \)-subscript on \( X, \beta \) is deliberate; the same coefficients and the same explanatory variables can occur in different equations. The earlier assumptions about the \( e_i \) are retained.

With a nonlinear model, \( f(\beta) \propto constant \) may no longer be suitable as a noninformative prior; consideration needs to be given to the type of nonlinear function and to whether particular values for some parameters need to be excluded. Thus, we give results for a general prior on \( \beta \), denoted by \( f(\beta) \). We retain the noninformative prior \( f(\Sigma) \propto |\Sigma|^{-1/2} \), and assume a priori independence of \( \beta \) and \( \Sigma \). Thus, the prior p.d.f. is given by

\[ f(\beta, \Sigma) \propto |\Sigma|^{-(M+1)/2} f(\beta) \]  

The likelihood function can be written as

\[ f(y|\beta, \Sigma) = (2\pi)^{-MT/2} |\Sigma|^{-T/2} \times \exp\left(-\frac{1}{2}(y - h(X, \beta))^\top \Sigma^{-1} (y - h(X, \beta))\right) \]

\[ = (2\pi)^{-MT/2} |\Sigma|^{-T/2} \exp\left(-\frac{1}{2} \text{tr}(A \Sigma^{-1})\right) \]  

where \( y = (y_1, y_2, \ldots, y_M)^\top \), \( h = (h_1, h_2, \ldots, h_M)^\top \), and now, \( A \) is an \((M \times M)\) matrix with the \((i,j)\)th element given by

\[ [A]_{ij} = [y_i - h_i(X, \beta)] [y_j - h_j(X, \beta)] \]  

The joint posterior p.d.f. for \((\beta, \Sigma)\) is

\[ f(\beta, \Sigma | y) \propto f(\beta) |\Sigma|^{-(T+M+1)/2} \exp\left(-\frac{1}{2} \text{tr}(A \Sigma^{-1})\right) \]  

and, integrating out \( \Sigma \), the marginal posterior p.d.f. for \( \beta \) is

\[ f(\beta | y) \propto f(\beta) |A|^{-T/2} \]  

Thus, the posterior for \( \beta \) in the nonlinear SUR model involves the same determinant of sums of squares and cross-products of residuals as it does in the linear model. A more general prior has been added. (Of course, it could also have been included in the linear model.)
The Metropolis–Hastings algorithm described in Section 3.7 can be readily applied to the posterior p.d.f. in Eq. (36). Because the earlier results on conditional posterior p.d.f.’s for $\beta$ and the $\beta_i$ no longer hold, the draws need to be used directly to estimate posterior p.d.f’s and their moments.

5 IMPOSING LINEAR EQUALITY RESTRICTIONS

Economic applications of SUR models frequently involve linear restrictions on the coefficients. For example, the same coefficient may appear in more than one equation, and the Slutsky symmetry conditions in demand models lead to cross-equation restrictions, or one might want to hypothesize that all equations have the same coefficients vector. Under the existence of cross-equation linear restrictions, the Gibbs sampler using $\beta$ and $\Sigma$, and the Metropolis–Hastings algorithm, can still be used. However, the Gibbs sampler involving only $\beta_i$ is no longer applicable. If the restrictions are all within equation restrictions, all three algorithms are possible.

Suppose a set of $J$ linear restrictions is written as

$$R\beta = (R_1, R_2)(\eta, \gamma) = r \quad (37)$$

where $R_1$ is $(J \times J)$ and nonsingular, $R_2$ is $(J \times (K-J)]$, and $\eta$ and $\gamma$ are $J$ and $(K-J)$ dimensional subvectors of $\beta$, respectively. To make this partition, it may be necessary to reorder the elements in $\beta$. Correspondingly, we can reorder the columns of $X$ and partition it so that the linear SUR model can be written as

$$y = X\beta + e = (X_1, X_2)(\eta, \gamma) + e \quad (38)$$

This reordering may destroy the block-diagonal properties of $X$. From Eq. (37), we can solve for $\eta$ as

$$\eta = R_1^{-1}(r - R_2\gamma) \quad (39)$$

Substituting Eq. (39) into Eq. (38) and rearranging yields:

$$y - X_1R_1^{-1}r = (X_2 - X_1R_1^{-1}R_2)\gamma + e$$

or

$$z = Zy + e \quad (40)$$
where \( z = y - X_1R_1^{-1}r \) and \( Z = X_2 - X_1R_1^{-1}R_2 \) represent new sets of “observations.”

In general, \( Z \) and \( \gamma \) can no longer be partitioned unambiguously into \( M \) separate equations. However, the stochastic properties of \( e \) remain the same. Thus, all the results in Sections 2 and 3 that did not rely on partitioning of \( X \) and \( \beta \) can still be applied to the model in Eq. (40). In particular, a Gibbs sampler can be used to draw \( \gamma \) and \( \Sigma \), and the Metropolis–Hastings algorithm can be used to draw \( \gamma \), from their respective posterior p.d.f.’s.

### 6 IMPOSING INEQUALITY RESTRICTIONS

Possible inequality restrictions on the coefficients range from simple ones such as a sign restriction on a single coefficient to more complex ones such as enforcing the eigenvalues of a matrix of coefficients to be nonpositive. Letting the feasible region defined by the inequality constraints be denoted by \( S \), and defining the indicator function:

\[
I_S(\beta) = \begin{cases} 
1 & \text{for } \beta \in S \\
0 & \text{for } \beta \notin S
\end{cases}
\]

the inequality restrictions can be accommodated by setting up the otherwise noninformative prior p.d.f.:

\[
f(\beta, \Sigma) \propto |\Sigma|^{-(M+1)/2} I_S(\beta)
\]

Using Bayes’ theorem to combine this prior with the likelihood function in Eq. (5), we obtain the joint posterior p.d.f.:

\[
f(\beta, \Sigma | y) \propto f(y | \beta, \Sigma) f(\beta, \Sigma)
= |\Sigma|^{-(T+M+1)/2} \exp\left(-\frac{1}{2} (y - X\hat{\beta})' (\Sigma^{-1} \otimes I_T) (y - X\hat{\beta}) \right) I_S(\beta)
\]

From this result we can derive the following conditional and marginal posterior p.d.f.’s.

The conditional posterior p.d.f. for \((\beta | \Sigma)\) is the truncated multivariate normal distribution:

\[
f(\beta | \Sigma, y) \propto \exp\left(-\frac{1}{2} (\beta - \hat{\beta})' X (\Sigma^{-1} \otimes I_T) X (\beta - \hat{\beta}) \right) I_S(\beta)
\]
The conditional posterior p.d.f. for \((\Sigma | y)\) is the same inverted Wishart distribution as was given in Eq. (20). The marginal posterior p.d.f. for \(\beta\) is

\[
f(\beta | y) \propto |A|^{-T/2} I_2(\beta)
\]

(45)

The posterior p.d.f. for \(\beta_1\) conditional on the remaining \(\beta_i\) is the truncated multivariate \(t\)-distribution:

\[
f(\beta_1 | y, \beta_2, \beta_3, \ldots, \beta_M) \propto \left[ v_1 + \frac{(\beta_1 - \tilde{\beta}_1)^T Q_{(1)} X_1 (\beta_1 - \tilde{\beta}_1)}{\tilde{x}_1^2} \right]^{-(K_1 + v_1)/2} I_2(\beta)
\]

(46)

Of interest is how best to use these p.d.f.'s to draw observations on \(\beta\), and possibly \(\Sigma\), from their respective posterior p.d.f.'s. The conditional posterior p.d.f.'s for \((\beta | \Sigma)\) and \((\Sigma | \beta)\) can be used within a Gibbs sampler providing that the inequality restrictions are sufficiently mild for a simple acceptance–rejection algorithm to be practical when sampling from the truncated multivariate normal distribution. By a “simple acceptance–rejection algorithm,” we mean that a draw is made from a nontruncated multivariate normal distribution and, if it lies outside the feasible region, it is discarded and replaced by another draw. This procedure will not be practical if the probability of obtaining a draw within the feasible region is small, which will almost always be the case if the number of inequality restrictions is moderate to large. Thus, we are using the term “mild” inequality restrictions to describe a situation where the maximum number of draws necessary before a feasible draw is obtained is not excessive.

If the inequality restrictions are not mild, then a Metropolis–Hastings algorithm can be employed. In the steps we described in Section 3, if a candidate value \(\tilde{\beta}^e\) is infeasible, then \(r = 0\), and the retained draw is automatically the last accepted feasible draw. That is, \(\beta^{(r+1)} = \beta^{(r)}\).

If the inequality restrictions are not mild, but are linear, then using a Gibbs sampler on subcomponents of \(\beta\) might prove successful. For example, using the truncated multivariate \(t\)-distributions for each of the \(\beta_i\), as specified in Eq. (46), could be useful. Also, within different contexts, sampling from truncated multivariate \(t\) and multivariate normal distributions has been broken down into sampling from univariate conditional distributions by Geweke (1991) and Hajivassiliou and McFadden (1990). See also the Appendix.
7 THREE APPLICATIONS

7.1 Wheat Yield

In Griffiths et al. (2001) the following model was used for predicting wheat yield in five Western Australian shires:

\[
Y_t = \beta_1 + \beta_2 t + \beta_3 t^2 + \beta_4 t^3 + \beta_5 G_t + \beta_6 G_t^2 \\
+ \beta_7 D_t + \beta_8 D_t^2 + \beta_9 F_t + \beta_{10} F_t^2 + e_t
\]

Yield \((Y_t)\) depends on a cubic time trend to capture technological change and on quadratic functions of rainfall during the germination period \((G_t)\), the development period \((D_t)\), and the flowering period \((F_t)\). The rainfalls are measured relative to their sample means. Inequality restrictions are imposed to ensure that the response of yield to rainfall, at average rainfall, is positive. That is, for germination rainfall, for example, \(\partial Y / \partial G = \beta_5 + 2\beta_6 > 0\). Thus, the feasible region for this example is

\[
S(\beta) = \{ \beta | \beta_5 + 2\beta_6 > 0, \beta_7 + 2\beta_8 > 0, \beta_9 + 2\beta_{10} > 0 \}
\]

Although Griffiths et al. (2001) used separate single equation estimation for the five shires and focused on several forecasting issues, investigation within a five-equation SUR model has started. Given that the inequality restrictions within each equation are relatively mild, but in total they are not, a Gibbs sampler using the truncated \(t\) densities in Eq. (46) seems a profitable direction to follow. Also, some preliminary work involving the Metropolis–Hastings algorithm on the complete \(\beta\) vector has proved effective.

7.2 Cost and Share Equations

In a second application, a translog cost function (constant returns to scale) and cost-share equations for merino woolgrowers (310 observations over 23 years) was estimated by Griffiths et al. (2000) using, as inputs, land, capital, livestock, and other factors. In the equations that follow \(c\) is cost, \(q\) is output, the \(w_i\) are input prices, and the \(S_i\) are input shares:

\[
\log\left( \frac{c}{q} \right) = \beta_0 + \beta_7 T + \sum_{j=1}^{4} \beta_{1j} \log(w_j) + 0.5 \sum_{i=1}^{4} \sum_{j=1}^{4} \beta_{ij} \log(w_i) \log(w_j) + e_1
\]

\[
S_i = \beta_j + \sum_{j=1}^{4} \beta_{ij} \log(w_j) + e_i \quad i = 2, 3, 4
\]

This SUR model has the following characteristics:
1. The equations are linear.
2. There are a number of linear equality restrictions that need to be imposed. Specifically, the $\beta_{ij}$s in the cost function are equal to the $\beta_{ij}$s in the share equations, and, furthermore, to satisfy homogeneity and symmetry, we require that

$$\sum_{i=1}^{4} \beta_i = 1 \quad \sum_{j=1}^{4} \beta_{ij} = 0 \quad \beta_{ij} = \beta_{ji}$$

3. Inequality restrictions are required for the functions to satisfy concavity and monotonicity. These restrictions are:
   - Monotonicity $0 < S_i < 1$
   - Concavity $B - S + ss'$ is negative semidefinite where

$$B = \begin{bmatrix}
\beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\
\beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \\
\beta_{31} & \beta_{32} & \beta_{33} & \beta_{34} \\
\beta_{41} & \beta_{42} & \beta_{43} & \beta_{44}
\end{bmatrix}$$

$$S = \begin{bmatrix}
S_1 \\
S_2 \\
S_3 \\
S_4
\end{bmatrix} \quad s = \begin{bmatrix}
S_1 \\
S_2 \\
S_3 \\
S_4
\end{bmatrix}$$

Note that $B - S + ss'$ is negative semidefinite if and only if its largest eigenvalue is nonpositive.

Since $S_i$ depend on the input prices, a decision concerning the input prices at which $S_i$ are evaluated, and the inequality restrictions imposed, needs to be made. The inequality restrictions were imposed at average input prices for each of the 23 years.

4. Given the severe inequality restrictions that were imposed, the Metropolis–Hastings algorithm was used.

5. The quantities of interest are nonlinear functions of the parameters. They are the elasticities of substitution:

$$\sigma_{ij} = \frac{\beta_{ij}}{SS_j} + 1 \quad i \neq j$$
and input demand elasticities:
\[ \eta_{ii} = \frac{\beta_{ii}}{S_i} + S_i - 1 \]
\[ \eta_{ij} = \frac{\beta_{ij}}{S_i} + S_j \]

### 7.3 Expenditure Functions

Our third example involves two expenditure functions estimated from a sample of 1834 Bangkok households, and deflated by an “equivalence scale” measure of household size (Griffiths and Chotikapanich, 1997). For the \( th \) observation, the functions are

\[
\begin{align*}
    w_{jt} & = \frac{\alpha_1 m_{1t}}{x_t} + \frac{\beta_1 m_{1t}(x_t - \alpha_1 m_{1t} - \alpha_2 m_{2t} - \alpha_3 m_{3t})}{x_t(\beta_1 m_{1t} + \beta_2 m_{2t} + \beta_3 m_{3t})} + e_{1t} \\
    w_{2t} & = \frac{\alpha_2 m_{2t}}{x_t} + \frac{\beta_2 m_{2t}(x_t - \alpha_1 m_{1t} - \alpha_2 m_{2t} - \alpha_3 m_{3t})}{x_t(\beta_1 m_{1t} + \beta_2 m_{2t} + \beta_3 m_{3t})} + e_{2t} \\
    m_{1t} & = 1 + \delta_{21} n_{2t} + \delta_{31} n_{3t} \\
    m_{2t} & = 1 + \delta_{22} n_{2t} + \delta_{32} n_{3t} \\
    m_{3t} & = 1 + \delta_{23} n_{2t} + \delta_{33} n_{3t}
\end{align*}
\]

where

\[ w_{jt} = \text{expenditure proportion for commodity } j, \]
\[ x_t = \text{total expenditure}, \]
\[ m_{jt} = \text{equivalence scale for commodity } j, \]
\[ n_{2t} = \text{number of extra adults (each household has at least one adult)}, \]
\[ n_{3t} = \text{number of children}. \]

The unknown parameters are

\[ (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \delta_{31}, \delta_{32}, \delta_{33}, \delta_{33}) \]

This SUR model has the following characteristics:

1. The equations are nonlinear in the parameters.
2. A number of inequality restrictions were imposed, namely,
   \[ 0 < \beta_1, \beta_2 < 1 \] additional expenditure from a one-unit increase in supernumerary income must lie between zero and one.
   \[ 0 \leq \delta_{31} \leq \delta_{32} \leq 1 \] expenditure requirements for extra adults are less than those for the first adult but greater than those for children.

\[ \alpha_i < \min_i \left\{ \frac{e_{it}}{1 + \delta_{2i} n_{2t} + \delta_{3i} n_{3t}} \right\} \quad i = 1, 2 \]
The smallest level of consumption in the sample must be greater than subsistence expenditure, a constraint from the utility function.

3. Given the nonlinear equations and the inequality constraints, the Metropolis–Hastings algorithm was used.

4. Two nonlinear functions of the parameters are of interest. They are the general scale or “household size”:

\[
m_0 = \frac{x \sum \beta_k m_k}{x - \sum \alpha_k m_k + \sum \alpha_k \sum \beta_k m_k}
\]

and the elasticities. Expressions for the latter can be found in Griffiths and Chotikapanich (1997).

8 FORECASTING

Suppose that we are interested in forecasting dependent variable values in the next period. The shire-level wheat yield application in the Section 7.1 is an example of where such a forecast would be of interest. In that case the objective is to forecast yield for each of the five shires. Since the yields are correlated via the stochastic assumptions of the SUR model, a joint forecast is appropriate. We can write the next period’s observation as

\[
y_s = X_s \beta + e_s
\]

where \( y_s \) is an \( M \)-dimensional vector, \( X_s \) is an \((M \times K)\) block diagonal matrix with the \( i \)th block being a \((1 \times K_i)\) row vector containing next period’s explanatory variables for the \( i \)th equation:

\[
X_s = \begin{bmatrix} x_{1s} & x_{2s} & \cdots & x_{Ms} \end{bmatrix}
\]

and \( e_s \sim N(0, \Sigma) \) is next period’s \((M \times 1)\) random error vector. The conventional Bayesian forecasting tool is the predictive p.d.f. \( f(y_s | y) \). Graphing marginal predictive p.d.f.’s from this density function, and computing its means, standard deviations, and probabilities of interest are the standard ways of reporting results.
The procedure for deriving the predictive p.d.f. is to begin with the joint p.d.f. \( f(y, \beta, \Sigma | y) \) and then to integrate out \( \Sigma \) and \( \beta \), either analytically or via a numerical sampling algorithm. Now,

\[
f(y, \beta, \Sigma | y) = \frac{1}{(2\pi)^{M/2}|\Sigma|^{-1/2}} \exp\left[-\frac{1}{2}(y - X\beta)'\Sigma^{-1}(y - X\beta)\right]
\]

\[
\propto |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}\text{tr}(A_*^{-1})\right]
\]

(51)

where \( A_* = [y - X\beta][y - X\beta]' \). Thus, using the posterior p.d.f. in Eq. (12) (no inequality restrictions), we have

\[
f(y, \Sigma | y) = f(y, \beta, \Sigma | y) f(\beta, \Sigma | y)
\]

\[
\propto |\Sigma|^{-(T+M+2)/2} \exp\left[-\frac{1}{2}\text{tr}((A + A_*)\Sigma^{-1})\right]
\]

Using properties of the inverted Wishart distribution to integrate out \( \Sigma \) yields:

\[
f(y, \beta | y) = \int f(y, \beta, \Sigma | y) d\Sigma \propto |A + A_*|^{-(T+1)/2}
\]

(52)

Because analytical integration of \( \beta \) out of Eq. (52) is not possible, we consider the conditional predictive p.d.f. \( f(y_* | \beta, y) \). It turns out that this p.d.f. is a multivariate student \( t \). Thus, \( f(y_* | \beta, y) \) and its moments can be estimated by averaging quantities from \( f(y_* | \beta, y) \) over draws of \( \beta \) obtained using one of the MCMC algorithms described earlier.

To establish that \( f(y_* | \beta, y) \) is a multivariate \( t \)-distribution, we first note that (see, e.g., Dhrymes 1978, p. 458)

\[
|A + A_*| = |A|(1 + (y_* - X\beta)'A^{-1}(y_* - X\beta))
\]

(53)

Thus,

\[
f(y_* | \beta, y) \propto [1 + (y_* - X\beta)'A^{-1}(y_* - X\beta)]^{-(T+1)/2}
\]

\[
\propto \left[ v_* + (y_* - X\beta)'\left(\frac{A}{v_*}\right)^{-1}(y_* - X\beta)\right]^{-(M+v_*)/2}
\]

(54)

where \( v_* = T - M + 1 \). Equation (54) is a multivariate \( t \)-distribution with mean:

\[
E(y_* | \beta, y) = X\beta
\]

(55)

covariance matrix:

\[
V(y_* | \beta, y) = \frac{A}{v_* - 2}
\]

(56)
and degrees of freedom \( v_u \). Given draws \( \beta^{(\ell)}, \ell = 1, 2, \ldots, N \) from an MCMC algorithm, one can average the quantities in Eqs. (54)–(56) over these draws to estimate the required marginal predictive p.d.f.’s and their moments. Marginal univariate \( t \) distributions from Eq. (54) are averaged and the formulas are analogous to those in Eqs. (22), (26), and (27) except, of course, that our random variable of interest is now an element of \( y_u \), say \( y_{ui} \), not \( \beta_{ik} \).

Percy (1992) describes an alternative Gibbs sampling approach where \( y_u, \beta \), and \( \Sigma \) are recursively generated from their respective conditional p.d.f.’s. With our approach, it is not necessary to generate draws on \( y_u \). Also, because we have derived the predictive p.d.f. conditional on \( \beta \), the introduction of inequality restrictions on \( \beta \) does not change the analysis. The range of values of \( \beta \) over which averaging takes place is restricted, but that is accommodated by the way in which \( \beta \) is drawn, and the result in Eq. (54) still holds.

An interesting extension, and one that is of concern to Griffiths et al. (2001), is capturing the extra uncertainty created by not knowing the value of one or more regressors in \( X^* \). We have this problem if a wheat-yield forecast is made prior to all rainfalls having been observed. The effect can be captured by modeling rainfall and averaging the predictive p.d.f. for yield conditional on rainfall over rainfalls draws made from its predictive p.d.f.

9 SOME EXTENSIONS

Consider estimating \( \beta \) in the SUR model when there are missing observations on one or more of the dependent variables. This problem was considered in the context of expenditure functions by Supat (1996). For the moment, assume that the observations are truly missing and that they are missing at random; they are not zeros created by negative values of an unobserved latent variable, as in the case with the Tobit model. Writing \( y^O \) to denote observed components and \( y^U \) to denote unobserved components, estimation can proceed within a Gibbs sampling framework using the conditional posterior p.d.f.’s \( f(\beta|\Sigma, y^O, y^U) \), \( f(\Sigma|\beta, y^O, y^U) \), and \( f(y^U|\beta, \Sigma, y^O) \). The conditional posterior p.d.f.’s for \( \beta \) and \( \Sigma \) are the normal and inverted Wishart p.d.f.’s given in Eqs. (14) and (20). To investigate how to draw observations from \( f(y^U|\beta, \Sigma, y^O) \), we write the \((M \times 1)\) th observation \( y_{(t)} \) as

\[
y_{(t)} = X_{(t)}\beta + e_{(t)}
\]

The subscript \( t \) has been placed in parentheses to distinguish the \((M \times 1)\) th observation on all equations \( y_{(t)} \) from the \((T \times 1)\) observations on the \( t \)th equation \( y_t \). The structure of \( X_{(t)} \) is similar to that of \( X^* \) defined in Eq. (50). We wish to consider Eq. (57) for all values of \( t \) where \( y_{(t)} \) has one or
more unobserved components. Reordering the elements if necessary, we can partition Eq. (57) as

\[
\begin{bmatrix}
  y_U(t) \\
  y_O(t)
\end{bmatrix} = \begin{bmatrix}
  X_U(t) \\
  X_O(t)
\end{bmatrix} \beta + \begin{bmatrix}
  e_U(t) \\
  e_O(t)
\end{bmatrix}
\]

where we write

\[
E[e(t)\mid e_0] = \begin{bmatrix}
  \Sigma^{UU} & \Sigma^{UO} \\
  \Sigma^{OU} & \Sigma^{OO}
\end{bmatrix}
\]

The conditional posterior p.d.f. \( f(y_U(t) \mid \beta, \Sigma, y_O(t)) \) is a multivariate normal distribution with mean:

\[
E(y_U(t) \mid \beta, \Sigma, y_O(t)) = X_O(t)\beta + \Sigma^{UO}\Sigma^{OO}^{-1}(y_O(t) - X_O(t)\beta)
\]

and covariance matrix:

\[
V(y_U(t) \mid \beta, \Sigma, y_O(t)) = \Sigma^{UU} - \Sigma^{UO}\Sigma^{OO}^{-1}\Sigma^{OU}
\]

Furthermore, \( (y_U(t) \mid \beta, \Sigma, y_O(t)), t = 1, 2, \ldots, T \) are independent. Thus, for generating \( y_U(t) \) within the Gibbs sampler, we use the conditional normal distributions defined by Eqs. (60) and (61) for all observations where an unobserved component is present.

Suppose, now, that the unobserved components represent negative values of a Tobit-type latent variable. In this case we have the additional posterior information that the elements of \( y_U(t) \) are negative. The conditional posterior p.d.f. for \( (y_U(t) \mid \beta, \Sigma, y_O(t)) \) becomes a truncated (multivariate) normal distribution with a truncation that forces \( y_U(t) \) to be negative. Its location vector and scale matrix (no longer the mean and covariance matrix) are given in Eqs. (60) and (61). A convenient algorithm for drawing from this truncated normal distribution is described in the Appendix.

For extensions into Probit models, see Geweke et al. (1997) and references cited therein. The literature on simultaneous equation models with Tobit and Probit variables can be accessed through Li (1998). Sets of SUR expenditure functions with a common parameter and with unobserved expenditures that result from infrequency of purchase are considered by Griffiths and Valenzuela (1998). Smith and Kohn (2000) study Bayesian estimation of nonparametric SURs.

10 CONCLUDING REMARKS

With the recent explosion of literature on MCMC techniques, Bayesian inference in the SUR model has become a practical reality. However, it is the author’s view that, prior to the writing of this chapter, the relevant results...
have not been collected and summarized in a form convenient for applied researchers to implement. It is my hope that this chapter will facilitate and motivate many more applications of Bayesian inference in the SUR model.

ACKNOWLEDGMENTS

I am grateful to Michael Chua for research assistance, and to Denzil Fiebig and Chris O’Donnell for comments on an earlier draft.

REFERENCES


Bayesian Inference in Regressions Model


**APPENDIX: DRAWING RANDOM VARIABLES AND VECTORS**

**A.1 Multivariate Normal Distribution**

To draw a vector $y$ from a $N(\mu, \Sigma)$ distribution:

1. Compute the Cholesky decomposition $H$ such that $HH^T = \Sigma$
2. Generate $z$ from $N(0, I)$
3. Calculate $y = \mu + Hz$

**A.2 Multivariate t-Distribution**

Consider the multivariate $k$-dimensional $t$-distribution with p.d.f.:

$$f(x | \mu, V) \propto [v + (x - \mu)'V^{-1}(x - \mu)]^{-(k+1)/2}$$

It has $v$ degrees of freedom, mean $\mu$, and covariance matrix ($v/(v-2))V$. (Assume $v > 2$.) To draw a vector $x$ from this p.d.f.:

1. Compute the Cholesky decomposition $H$ such that $HH^T = V$
2. Generate the $(k \times 1)$ vector $z_1$ from $N(0, I_k)$
3. Generate the $(v \times 1)$ vector $z_2$ from $N(0, I)$
4. Calculate $x = \mu + Hz_1/\sqrt{z_2^Tz_2/v}$
A.3 Inverted Wishart Distribution

Let $\Sigma$ have an $m$-dimensional inverted Wishart distribution with parameter matrix $S$ and degrees of freedom $v$. It has p.d.f.:

$$f(\Sigma | S) \propto |\Sigma|^{-(v+m+1)/2} \exp\left(\frac{-1}{2} \text{tr}(S\Sigma^{-1})\right)$$

To draw observations on $\Sigma$:

1. Compute the Cholesky decomposition $H$ such that $HH^T = S^{-1}$.
2. Draw independent ($m \times 1$) normal random vectors $z_1, z_2, \ldots, z_v$ from $N(0, I_m)$.
3. Calculate $\Sigma = (H \sum_{i=1}^v z_i z_i^T H^T)^{-1}$.

A.4 Univariate Truncated Normal Distribution

Suppose that $x$ is a truncated normal random variable with location $\mu$, scale $\sigma$, and truncation $a < x < b$. To draw $x$:

1. Draw a uniform (0, 1) random variable $U$.
2. Calculate:

$$x = \mu + \sigma \Phi^{-1}\left[\frac{a - \mu}{\sigma}\right] + U\left(\Phi\left(\frac{b - u}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)\right) \quad (A.1)$$

where $\Phi$ is the standard normal cumulative distribution function.

A.5 Multivariate Truncated Normal Distribution

Suppose that $x$ is an $m$-dimensional multivariate truncated normal distribution such that $a_1 < x_1 < b_1$, $a_2 < x_2 < b_2$, $\ldots$, $a_m < x_m < b_m$.

1. Use Eq. (A.1) to draw $x_1$.
2. Find the location and scale parameters for the truncated conditional normal distribution ($x_2 | x_1$) conditional on $x_1$ drawn in step 1.
3. Apply Eq. (A.1) to the distribution ($x_2 | x_1$).
4. Find location and scale parameters for the distribution of ($x_3 | x_2, x_1$) conditional on the draws made in steps 1 and 3.
5. Apply Eq. (A.1) to the distribution ($x_3 | x_2, x_1$).
6. And so on.
10

Computationally Intensive Methods for Deriving Optimal Trimming Parameters

Marco van Akkeren
University of California at Berkeley, Berkeley, California, U.S.A.

1 INTRODUCTION

The information recovery process envisions sample data that are often limited, partial, or ill-conditioned, and the corresponding statistical models that form the basis for estimation and inference may in some cases be underdetermined or ill-posed. The ill-posed aspect may arise because nonstationary or other model specification reasons may cause the number of parameters to exceed the number of data points or the moment conditions to be less than or greater than the number of unknown parameters. In addition, the design matrix implied by nature or society may be ill-conditioned. Consequently, if traditional estimation procedures are used, the solution may be undefined or the estimates that are obtained may be highly unstable giving rise to high variance and low precision of the recovered parameters.

To have a formal basis for the analysis of these problems, consider the following probability model: suppose we make \( n \) observations on \( k \) random variables \( y_i, x_{2i}, x_{3i}, \ldots, x_{ki} \). We can specify a relationship between the response variable \( y_i \) and the conditioning variables \( x_{2i}, x_{3i}, \ldots, x_{ki} \).
using the statistical model:
\[ y = X\beta + e \]  
where \( x_{ij} = i \) to indicate inclusion of the intercept term, \( y = (y_1, \ldots, y_n) \) is the vector of observed sample values, \( X \) is a known \((n \times k)\) nonstochastic design matrix, and \( \beta \in \mathcal{B} \subset \mathbb{R}^k \) represents the vector of unknown parameters. The \( n \)-dimensional noise vector \( e \) is assumed to represent a drawing from an unknown multivariate distribution with zero mean and covariance matrix \( \Sigma_e = \sigma^2 I_n \). This is the usual linear statistical model. For notation purposes, we adhere to the convention where bold capital letters refer to matrices and bold lower case letters refer to column vectors.

Given the linear statistical model, one plausible criterion is to find an estimator \( \hat{\delta}(y) \in \mathcal{D} \) of the unknown parameter vector \( \beta \in \mathcal{B} \) that yields small expected squared error loss (SEL) relative to conventional estimators:
\[ R(\hat{\delta}(y)|\beta) = \mathbb{E}_\beta \| \hat{\delta}(y) - \beta \|^2 \]  
If \( \beta \in \mathcal{B} \subset \mathbb{R}^k \) and under the SEL measure, the least squares (LS) estimator \( \hat{\beta}(y) \) is best linear unbiased with constant risk equal to \( \sigma^2 \text{tr}(X'X)^{-1} \). If \( e \) is a multivariate normal distributed random variable then the maximum likelihood estimator \( \hat{\delta}^0(y) \) of \( \beta \) equals the LS estimator \( \hat{\beta}(y) \):
\[ \hat{\delta}^0(y) = \hat{\beta}(y) = (X'X)^{-1}Xy \sim N_k(\beta, \sigma^2 (X'X)^{-1}) \]  
where \( \hat{\delta}^0(y) \) is best unbiased and minimax under SEL and which is denoted by the notation \( \hat{\delta}^*(y) \):
\[ \sup_{\beta \in \mathcal{B}} R(\hat{\delta}^0(y)|\beta) = \min_{\hat{\delta}(y) \in \mathcal{D}} \sup_{\beta \in \mathcal{B}} R(\hat{\delta}(y)|\beta) \]  
The linear statistical model formulation Eq. (1), and the corresponding estimation rules, \( \hat{\beta}(y) \) and \( \hat{\delta}^0(y) \), form the basis for a range of statistical models involving sample observations from both discrete and continuous distributions.

In order to evaluate the performance of Eq. (3) more effectively, we should note that Stein (1955) and James and Stein (1961) showed the inadmissibility of \( \hat{\delta}^0(y) \) for \( k > 2 \) using the traditional Stein estimator:
\[ \hat{\delta}^*(b) = \left[ 1 - \frac{(k-2)\sigma^2}{\|b\|^2} \right] b \]  
which implies that \( R(\hat{\delta}^*(b)|\beta) \leq R(\hat{\delta}^0(b)|\beta) \) for all \( \beta \in \mathcal{B} \) and with strict inequality for at least one \( \beta \in \mathcal{B} \). Baranchik (1970) proved that the traditional Stein estimator itself is dominated by Stein’s positive-rule.
estimator:

\[ \delta^+(b) = \left[ 1 - \frac{(k-2)\sigma^2}{\|b\|^2} \right]_+ b \]  

(6)

where \([u]_+\) denotes \(\max\{0, u\}\), (Judge and Bock, 1978), which has been shown to be inadmissible by Efron and Morris (1973). Investigation of the above formulations shows that the Stein estimator shrinks each element of \(b\) by the same amount towards zero. This does not need to be the case. Suppose a priori information is available that reduces the parameter space and allows for specification of a constant target vector \(c \in \mathbb{R}^k\). The Stein’s modified estimator results:

\[ \delta^m(b,c) = \left[ 1 - \frac{(k-2)\sigma^2}{\|b-c\|^2} \right] (b - c) + c \]  

(7)

which shrinks \(b\) towards the target \(c\) and is itself also dominated by a positive-rule modified Stein estimator. It can be shown that if \(c\) is appropriately chosen, i.e., \(k^{-1}i\|b\|\) where \(i\) is a vector of ones, then \(\delta^m(b,c)\) is minimax for \(k > 3\). For the modified Stein estimator to be effective, sufficient a priori information must be available to reduce the parameter space and accurately conjecture the target vector. If \(b\) is close to \(c\) then significant improvement is made in \(R(\delta^m(b,c)|\beta)\), which dissipates quickly the larger is \(\|c-\beta^2\|\) or the farther is \(c\) from a neighborhood around \(\beta\). Consequently, the quality of nonsample information is of importance to the sampling performance of Stein-like estimators.

More recently, Zellner (1994) introduced the balanced loss function as a means to incorporate both goodness of fit and precision of estimation in the estimator performance criterion. The general form of this loss function is

\[ L_w(\delta(y)|\beta) = w\|y - X\delta(y)\|^2 + (1 - w)\|\delta(y) - \beta\|^2 \]  

(8)

where \(0 \leq w \leq 1\) is the relative weight associated with prediction. Dey et al. (1999) have shown that if the estimator of the location parameters, expressed as \(\delta(y) = b(y) + g(y)\) dominates the LS estimator \(b\) where \(b(y)\) is defined by Eq. (3) and \(g(y): \mathbb{R}^n \rightarrow \mathbb{R}^k\), then \(\delta_* = b(y) + (1-w)g(y)\) dominates \(b\) under \(L_w(\delta|\beta)\) for all \(w\). The Stein estimator consequently may be represented in this form and hence can be shown also to dominate the LS estimator under Zellner’s balanced loss function. It is under this criterion that van Akkeren (1999) and van Akkeren, Judge, and Mittelhammer (2002) developed the data-based information theoretic (DBIT) estimator, which has attractive sampling properties in the case of
ill-conditioned linear statistical model. With these concepts in mind, Section 2 develops the formulations for the DBIT estimator along with the ridge and principal components estimators where the optimal tuning parameter is chosen to minimize an unbiased estimator of the risk.

2 INFORMATION PROCESSING RULES

Given the framework of ill-conditioned inverse problems, we first review the DBIT estimator and state the properties that characterize its sampling behavior. Sections 2.2 and 2.3 will describe the ridge and principal components estimators, which are contained in the more general class of method of regularization estimators. Our review will be fairly general and is intended to give an overview of these methods. For more specific details, the reader is encouraged to examine the original papers.

2.1 The DBIT Estimator

In developing an information theoretic alternative estimation procedure, we generalize the traditional estimating function assumption \( g(y, \beta) = X(y - X\beta) = 0 \) and make use of information about the unknown distribution and parameters in the form of the following estimating function:

\[
  h(y, \beta, e) = X'(y - X\beta - e) = 0
\]

or equivalently,

\[
  n^{-1}[X'y - X'X\beta - X'e] = 0
\]

The \( k \)-dimensional structural data constraint contains \( (k + n) \) unobservables, \( (\beta, e) \), and thus is ill-posed. Consequently, traditional matrix procedures can not yield a unique solution that satisfies the equation.

The structural data constraints, Eqs. (9) and (10), express one representation of the observable information that underlies the unknown and unobservable \( \beta \) and \( e \). Traditional information-processing rules are usually expressed as some function of the sample information either in a data or moment form and belong to the general class of estimators

\[
  \Psi = \{ \psi: \delta(y) = \sum_{j=1}^{n} (\psi_j \delta_j(y)) \}.
\]

For example, in the case of the LS estimator, we have \( \psi_j = (X'X)^{-1}x_j \), which reduces to \( \psi_j = x_j \) for the orthonormal linear model. Also, note that in the case of IV(\( h = k \)) we have \( \psi_j = (Z'Z)^{-1}z_j \) or \( \psi_j = [(X'Z)(Z'Z)^{-1}(X'Z)Z^{-1}z_j \) for IV(\( h > k \)). Consequently, all are some weighted sum of \( y \) values, and all belong to the same estimator class given appropriate choices of \( \psi_j \). To solve the problem, we seek estimates of \( \delta(y) \) and \( \gamma(y) \) that satisfy Eqs. (9) and (10). In particular, expressing the observables \( y \) and \( X \) in terms of deviations about the mean, we seek \( p \) and \( w \)
that satisfy
\[ n^{-1}[X'\hat{y} - X'(X' \odot P)y - X'(I_n \otimes y')w] = 0 \] (11)
such that \( \delta(y) = (X' \odot P)y \) with the \((k \times n)\) matrix \( P = [p_1, p_2, \ldots, p_k]' \) and \( \gamma(y) = (I_n \otimes y')w \), where \( w' = [w_1', w_2', \ldots, w_n'] \) and \( p' = [p_1', p_2', \ldots, p_k'] \). We make use of the Hadamard product \( \odot \) and let the weights \( p \) and \( w \) represent \((k + n)\) univariate probability distributions, one for each element of \( \delta \) and \( \gamma \). Thus, in our DBIT formulation, we apply the sample weights \( \pi \) marginally (coordinate wise) and specify a univariate probability distribution \( p_k \) on the sample observations for each of the estimates \( \delta_k(y) \). Correspondingly, for the estimates of the disturbance terms \( \gamma(y) \), a univariate probability distribution \( w_i \) on the sample observations is specified for each \( \gamma_i(y) \), implying each \( e_i \) is contained in the convex hull of the centered sample points. Consistent with moment information in the form of Eqs. 9–11, and the Kullback-Leibler (KL) pseudo-distance measure, we have the extremum problem proposed by van Akkeren and Judge (2001):

\[ \min_{p, w} I(p, q, w, u) = p' \log(p/q) + w' \log(w/u) \] (12)
subject to
\[ \frac{X'\hat{y}}{n} = (X' \odot P)y + \frac{(n^{-1})}{n} X'(I_n \otimes y')w \] (13)
\[ i_k = (I_k 
abla i')_p \] (14)
\[ i_n = (I_n 
abla i')_w \] (15)

where \( q \) and \( u \) are the reference distributions composed of uniform probabilities in the case of noninformative priors. In this constrained optimization problem, the objective is to recover the \( p \) and \( w \) and thereby derive the optimal estimates of \( \beta \) and \( e \).

Given the objective of minimizing KL information that is inherent to the definition of the estimator, there is a tendency to draw the noise component vector to the centroid of the polyhedron formed from the centered \( y \) values. That is, KL information, which is based on an extension of Laplace’s principle of insufficient reason or indifference, takes its global maximal value when all convexity weights are \( n^{-1} \), which would draw the error estimates to their central tendency of the zero vector. Thus, imposing convexity weights on the feasible space of errors would seem a natural restriction. In particular, without further information, errors are estimated to be equal to their expected values. Regarding the normalization of the weights to have unit sum, note that traditional estimators also impose restrictions on the weights applied to the \( y_i \) values when determining the feasible values of the errors. For example, in the least squares case, the
weights are defined by the positive semidefinite nature of the idempotent matrix, \((I - XX^{-1}X')\), that premultiplies \(y\) in the definition of the estimate of the error vector. In this case, each of the row sums of the idempotent matrix, which represents the sum of the weights applied to the \(y\) vector in defining each element of \(\gamma(y)\) within the LS estimator context, equals 1 when the data is centered as it is here.

In terms of the \(B\) space, the feasible values are defined by \((X' \otimes P)y\). Under the KL criterion function Eq. (12), and based on arguments analogous to those used in motivating the behavior of the error component definition, the objective relative to the \(\beta\) component has a tendency to draw all the convexity weights to \(n^{-1}\) and seek the weights that are most maximally uninformative and could have arisen the greatest number of ways. Thus, there is a tendency to draw the estimate of \(\beta\) towards \(n^{-1}X'y\). Note also that the row sums of the matrix, \((X'X)^{-1}X'\), weighting the \(y\) elements each equal zero. In comparison, the row sums of the weights applied to the \(y\) elements in defining the LS estimator of the \(\beta\) vector also equal zero when the data are centered.

Solving the constrained optimization problem defined by Eqs. (12)–(15) using the method of Lagrange multipliers, we obtain the optimal weights:

\[
\hat{p}_{km} = \frac{q_{km} \exp\left[\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{p} \hat{\lambda}_{ik} X_{ik} X_{km} y_{in}\right]}{\sum_{m=1}^{m} q_{km} \exp\left[\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{p} \hat{\lambda}_{ik} X_{ik} X_{km} y_{in}\right]} = \frac{q_{km} \exp(\cdot)}{\Omega_k(\hat{\lambda})} \quad \forall k, m
\]

\[
\hat{w}_{ij} = \frac{u_{ij} \exp\left[\frac{1}{n} \sum_{k=1}^{n} \hat{\lambda}_{ij} y_{ij}\right]}{\sum_{j=1}^{n} u_{ij} \exp\left[\frac{1}{n} \sum_{k=1}^{n} \hat{\lambda}_{ij} y_{ij}\right]} = \frac{u_{ij} \exp(\cdot)}{\Psi(\hat{\lambda})} \quad \forall i, j
\]

that are used to obtain the estimates \(\delta(y)\) and \(\gamma(y)\). The strict convexity of the objective function and linear constraints ensure that the optimal solution exists and is uniquely defined when weights are strictly positive. In addition, further examination of Eqs. (16) and (17) reveals that \(\hat{p}\) and \(\hat{w}\) are both continuous and monotonic functions of \(\hat{\lambda}\). These observations, in addition to certain regularity conditions, are used to show that the DBIT estimator is consistent and asymptotically normal. Consequently, the usual asymptotic tests may be applied for inference, and confidence regions can be estimated. Finite sampling properties of the DBIT estimator are investigated in van Akkeren, Judge and Mittelhammer (2002) and van Akkeren (1999) to show that DBIT estimates dominate...
the LS and traditional method of regularization (MOR) estimators under the SEL measure for different levels of ill-conditioning.

2.2 Method of Regularization

Traditionally, a restricted LS or penalized likelihood approach has been proposed to deal with ill-conditioned inverse problems. As such, a MOR estimator is used where the data are fit using a distance measure of goodness of fit $c(y, x; b)$ with penalty function $f(b)$ (Tikhonov and Arsenin, 1977; Titterington, 1985; O’Sullivan, 1986). It is clear that the properties of the resulting estimator can be derived using the theory of extremum estimation and estimating equations. Specifically, the ridge regressor $\delta'(y; \kappa)$ is defined as

$$ \delta'(y; \kappa) = \arg\min_{\beta \in \mathcal{B}} [c(y, X; \beta) + \kappa f(\beta)] $$

(18)

when $\kappa \to \infty$, the resulting ridge regressor is driven by the penalty function, while a value of $\kappa = 0$ implies an estimate based on goodness of fit. Hoerl and Kennard (1970a,b; 1976) propose the ridge estimator to counter problems of singularity and near singularity of $XX$ by adding matrix $\kappa I$, where $\kappa \in \mathbb{R}^+$ such that the inverse of the sum is well defined. It follows that the normal equations:

$$(XX + \kappa I)\delta = X'y$$

(19)

can be uniquely solved for $\delta'(y; \kappa)$. Taking expectations of $\delta'(y; \kappa)$ reveals the estimator is biased while $\text{Cov}(b(y)) - \text{Cov}(\delta'(y; \kappa))$ is shown to be positive semidefinite. The trade-off between bias and efficiency is employed to show that $R(\delta'(y; \kappa^*)|\beta) < R(b(y)|\beta)$ for optimal tuning parameter $\kappa^*$, which is dependent of $\beta$ and as such represents an impracticality to estimation.

Several ad hoc methods exist for selecting the optimal tuning parameter. One simple alternative sets $\kappa^* = \hat{s}^2/(k - 2)/||b(y)||^2$, which can be improved upon by applying an iterative ridge procedure (Hoerl and Kennard, 1976) that updates $\kappa^*$ until $\delta'(y; \kappa)$ stabilizes. More recent developments have led to $\kappa^*$, which is the argument that minimizes an unbiased estimator of the risk (Beran, 1998). If precision of estimation is considered, algebraic manipulation allows us to derive the following:

$$ R(\delta(y; \kappa)|\beta) = \mathbb{E}||\delta(y; \kappa) - \beta||^2 $$

$$ = \text{tr}[G_* (\sigma^2 X'X + \kappa^2 \beta\beta')G_*] $$

(20) (21)
\[
\begin{align*}
\sum_{i=1}^{k} \frac{\lambda_i}{(\lambda_i + \kappa)^2} &+ \frac{\kappa^2 \tilde{\beta}_i^2}{(\lambda_i + \kappa)^2} \\
&= \sum_{i=1}^{k} \frac{\sigma^2 \lambda_i + \kappa^2 \tilde{\beta}_i^2}{(\lambda_i + \kappa)^2}
\end{align*}
\]  

(22)

(23)

where \(\text{tr}[.]\) indicates the standard notation for the trace of [], \(G_e = (XX + \kappa I)^{-1}\), \(\tilde{\beta} = P^n \beta\), and the matrix \(P\) is obtained from spectral decomposition of \(XX = PA^P\) with \(\lambda_i\) being the \(i\)th characteristic root of \(XX\). The above formulation clearly contains two unknowns; the variance, \(\sigma^2\), and the transformed vector of parameters \(\tilde{\beta} = P^n \beta\). The variance parameter can be easily estimated using an unbiased estimator such as \(\sigma^2 = \frac{k}{n-k} \sum (y - Xb)^2\) whereas \(\tilde{\beta}_i^2\) may be estimated by \(\tilde{\beta}_i^2 = \frac{s^2}{\lambda_i}\). Hence, by substituting these estimates for the unknowns in Eq. (23) an unbiased estimator of the risk [Eq. (20)] is obtained:

\[
\hat{R}_1(\hat{\beta}, \hat{\sigma}^2) = \sum_{i=1}^{k} \frac{s^2 \lambda_i + \kappa^2 \tilde{\beta}_i^2}{(\lambda_i + \kappa)^2} \\
\]  

(24)

As stated, this empirical risk function addresses precision of estimation in terms of the SEL measure. If goodness of fit is considered, an alternative empirical prediction risk is obtained from the following algebraic manipulations:

\[
R_2(\delta(y; \kappa) | \beta) = \mathbb{E}[(Xy)^2 - X\beta)^2] \\
= \mathbb{E}[(Xy)^2 - X\beta)^2] \\
= \text{tr}[X\mathbb{E}(\delta(y; \kappa) - \beta)^2] \\
= \text{tr}[X\mathbb{E}(\sigma^2 X - \kappa^2 \beta \beta^T)X] \\
= \sum_{i=1}^{k} \frac{\sigma^2 \lambda_i + \kappa^2 \tilde{\beta}_i^2}{(\lambda_i + \kappa)^2}
\]  

(25)

(26)

(27)

(28)

(29)

The unknown quantities, \(\sigma^2\) and \(\tilde{\beta}_i^2\), in Eq. (29) can be replaced using the same unbiased estimators as before. To consider both prediction and precision of estimation in the estimation objective, a linear combination of \(\hat{R}_1\) and \(\hat{R}_2\) yields Zellner’s balanced loss function of the form:

\[
\hat{R}_w(\hat{\beta}, \hat{\sigma}^2) = w \hat{R}_1(\hat{\beta}, \hat{\sigma}^2) + (1 - w) \hat{R}_2(\hat{\beta}, \hat{\sigma}^2)
\]  

(30)

for a particular value of \(w \in [0,1]\). The value of \(\kappa^* \geq 0\) is selected to minimize the objective [Eq. (30)] and consequently is entirely dependent on
the data and not on nonsample information or a priori support space restrictions. While the approach for minimizing an unbiased estimator of the risk leads to a data-based procedure for selecting the optimal tuning parameter, the resulting estimator, $\hat{\delta}(y; k^*)$, is difficult to evaluate and the corresponding sampling properties are analytically intractable. Simulation and resampling methods may, however, represent a computationally convenient means to evaluate the point estimates.

2.3 Principal Components Estimator

In addition to ridge regression, one frequently used method for dealing with ill-conditioned data is to reduce the informational demands by considering a subset of the parameter space $B$. In principal components regression this amounts to removing the conditioning variables that contribute little to the explanatory power of the model under a specific criterion. Several procedures have been proposed for variable elimination (Hill et al., 1997; Fomby and Hill, 1978; Mundlak, 1981; Hill and Judge, 1987); however, these methods tend to lack rigorous theoretical justification and produce estimates with poor sampling properties.

Beran (1998) proposes an alternative formulation that addresses the identification problem of which specific column vectors of $X$ to remove. This method uses an approach that miminizes an unbiased estimator of the risk in a fashion similar to that of Section 2.2. The criterion under which to evaluate estimator performance simply becomes the squared error prediction loss (SEPL), or the balanced loss function $L_w(\hat{\delta}(y)/\hat{\beta})$, where $w = 1$. As a starting point for this model, denote by $P = (p_1, \ldots, p_k)$ that $(k \times k)$ orthogonal matrix whose columns $p_i$ are the eigenvectors of $XX^T$ numbered according to the magnitude of characteristic roots $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. The linear statistical model is then reparameterized to the following:

$$y = XPP^T \beta + e = \hat{X}\hat{\beta} + e$$  (31)

which is the principal components regression model. Here, the $i$th principal component is $\hat{x}_i = Xp_i$ and it follows that $\hat{x}_i^T\hat{x}_i = \lambda_i$. The principal components estimator is consequently formulated as $\hat{\beta}(y) = (\hat{X}^T\hat{X})^{-1}\hat{X}^Ty$ where the subscript indicates the first $k$ column vectors of $\hat{X}$. The concept behind this method becomes clear in that one would like to remove the $\hat{x}_i$ associated with a small characteristic root $\lambda_i$, yet the major question of how to determine the correct cut-off point remains.

In addressing this problem, we consider the canonical modulation estimator proposed by Beran (1998) and exploit the singular value
decomposition of $X = ULV'$ with the properties such that $U'U = I_k$, $VV' = I_k$ and $L = \text{diag}(l_1, l_2, \ldots, l_k)$ are the singular values. It can be shown using a few algebraic steps that the candidate estimator for $X\beta$ can be expressed as $X\hat{\beta} = UFU'y$ where the $(k \times k)$ matrix $F$ consists of a string of $k$ ones followed by $(k - k)$ zeros and all off-diagonal elements are zero as well. The prediction risk of this estimator where $Z = U'y$ and $\xi = LU\hat{\beta}$ is derived as follows:

$$ R(b_f(y)|\beta) = E\|UFU'y - ULV'\beta\|^2 $$

$$ = E\|FZ - \xi\|^2 $$

$$ = E\text{tr}([F(Z - \xi) - (I - F)\xi][F - (I - F)\xi']') $$

$$ = \text{tr}(F^2E(Z - \xi)(Z - \xi)' + (I - F)^2\xi\xi') $$

$$ = \text{tr}(\sigma^2F^2 + \xi^2(I - F)^2) $$

$$ = \sum_{i=1}^{k} [\sigma_i^2f_i^2 + \xi_i^2(1 - f_i)^2] $$

where the unknowns, $(\sigma^2, \xi^2)$, can be estimated by the unbiased estimators $s^2$ and $z_i^2 - s^2$ since it is observed that $E[z_i^2] = \sigma^2 + \xi_i^2$. Minimization of the empirical representation of Eq. (37) leads to the data-based estimate of $X\beta$, but derives an estimator functional form that is difficult to evaluate analytically for finite sample properties. Large sample properties of the resulting estimator represent one means to evaluate the estimates whereas resampling and Monte Carlo methods could provide a computationally simple alternative.

3 EXAMPLES

To give the discussed estimators operational content, this section investigates two sets of data that are frequently studied in the statistics literature because of particular problems that characterize the data. Estimates derived using the DBIT method are compared to usual LS results as well as those computed using principal components, ridge regression, and nonparametric smoothing techniques. One should keep in mind that while the discussion has focussed on small sample estimator performance in terms of precision of estimation and prediction, this section concentrates on goodness of fit as the criterion for comparison.
3.1 Motorcycle Data

The first example is a data set composed of 133 observations of acceleration against time for a simulated motorcycle accident that may be found in Silverman (1985) and is analyzed by Härdle (1991) and Venables and Ripley (1994). For each observation the measurement time is between 0 and 60 sec. Accordingly, we postulate that acceleration $y_i$ against time $t_i$ satisfies the trigonometric regression model:

$$ y_i = \beta_0 + \sum_{j=1}^{12} \left[ \gamma_j \cos(2\pi j t_i / 60) + \beta_j \sin(2\pi j t_i / 60) \right] + e_i \quad \text{for} \ 1 \leq i \leq 133 $$

(38)

Here, as usual, we assume the vector of unknown disturbances $e$ to represent drawings from an unknown multivariate distribution with mean $0$ and covariance matrix $\Sigma_e = \sigma^2 I$ in addition to standard assumptions of a nonstochastic design matrix $X$ and unknown vector of coefficients $\beta \in \mathbb{B}$ as stated in Section 1. The degree to which one column vector of $X$ lies in the subspace spanned by the other is reflected in the condition number $\kappa(X'X)$, which is defined as the ratio of the largest to the smallest characteristic root of $X'X$, and equals about 189.08 for this particular data set, implying that collinearity represents a substantial problem.

3.2 DBIT Regression Estimates

Given the large sample size of the data, $n = 133$, the limiting properties proposed by van Akkeren, Judge, and Mittelhammer (2001) suggest DBIT estimates should be fairly close to the true parameter $\beta_0$ and moreover, they should be normally distributed with close approximation to LS estimates. Solving the optimization problem, Eqs. (12)–(15), leads to a computational difficulty due to the many constraints and unknown variables of the model. As a practical solution, we propose reducing the number of targets for the $k$th parameter from $\{y_k y_{(i)}\}_{i=1}^n$ to $\{y_k y_{(j)}\}_{j=1}^n$, where $j$ refers to the number of ordered quantiles and $\gamma \in \mathbb{R}^+$ is a scale parameter that increases or tightens the distance between the empirical target points. For example, we may choose to select $j = 5$, leaving us with the minimum, maximum, the interquartile range, and the median. Likewise, we reduce the number of targets of the $i$th error term from $\{y_i - (1/n)\hat{y}_i\}_{i=1}^n$ to $\{y_i - (1/n)\hat{y}_i\}_{i=1}^n$, which sets the bounds of the support space as well as giving less weight to outlying target points.
Table 1 displays the DBIT regression estimates for $j = 11$ and 3 selected values of the scale parameter $\gamma$ with Fig. 1 showing the graphical representation of these estimates. It is observed that the data display an unusual nonlinear shape where the dispersion of points around the regression line varies from very wide to tight. Of the selected values of $\gamma$, it is clear that 0.1 sets supports that are too tight and hence not able to span the entire parameter space, resulting in the poor estimates found in column 2 of Table 1. A value of 1 or 10 for this parameter leads to more acceptable results with closer approximation to LS estimates. For parameter estimates under $\gamma = 1$ and 10, the variance estimates are accordingly 546.5530 and 545.9179 compared to LS estimate of 545.9045. Next, we compare the resulting regression fits with principal components and ridge regression techniques.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\gamma = 0.1$</th>
<th>$\gamma = 1$</th>
<th>$\gamma = 10$</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>1.27064</td>
<td>8.71496</td>
<td>9.95335</td>
<td>9.06142</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>7.06728</td>
<td>14.61594</td>
<td>15.40384</td>
<td>15.41838</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-8.04954</td>
<td>-23.49255</td>
<td>-23.17484</td>
<td>-23.11846</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>4.55280</td>
<td>14.32972</td>
<td>15.20501</td>
<td>15.20574</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.43147</td>
<td>-4.51787</td>
<td>-4.06511</td>
<td>-4.04139</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.90008</td>
<td>-2.65975</td>
<td>-2.16492</td>
<td>-2.15621</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-1.24315</td>
<td>-1.55327</td>
<td>-1.03614</td>
<td>-1.03692</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.58241</td>
<td>0.78432</td>
<td>1.19926</td>
<td>1.19661</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>0.21283</td>
<td>-0.62523</td>
<td>-0.32104</td>
<td>-0.31714</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.58363</td>
<td>-3.98569</td>
<td>-3.82590</td>
<td>-3.81345</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>-0.41283</td>
<td>-1.64735</td>
<td>-1.49423</td>
<td>-1.49148</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>-0.27299</td>
<td>-0.66754</td>
<td>-0.42911</td>
<td>-0.43592</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-6.33800</td>
<td>-28.36732</td>
<td>-29.22203</td>
<td>-29.09952</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>9.16865</td>
<td>30.81417</td>
<td>30.92620</td>
<td>30.86495</td>
</tr>
<tr>
<td>$\beta_{15}$</td>
<td>-3.80181</td>
<td>-12.58201</td>
<td>-12.90998</td>
<td>-12.89735</td>
</tr>
<tr>
<td>$\beta_{16}$</td>
<td>-2.68503</td>
<td>-6.30882</td>
<td>-6.40717</td>
<td>-6.41988</td>
</tr>
<tr>
<td>$\beta_{17}$</td>
<td>0.32144</td>
<td>3.01598</td>
<td>2.96387</td>
<td>2.95566</td>
</tr>
<tr>
<td>$\beta_{18}$</td>
<td>2.07148</td>
<td>0.74785</td>
<td>0.44230</td>
<td>0.45266</td>
</tr>
<tr>
<td>$\beta_{19}$</td>
<td>-0.00556</td>
<td>-0.71955</td>
<td>-0.87236</td>
<td>-0.87333</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>-2.25243</td>
<td>-2.48885</td>
<td>-2.58998</td>
<td>-2.60282</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>2.02036</td>
<td>-2.14160</td>
<td>-2.59033</td>
<td>-2.57853</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>0.64224</td>
<td>2.80111</td>
<td>2.81314</td>
<td>2.81422</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>-1.09331</td>
<td>-2.00492</td>
<td>-2.02609</td>
<td>-2.01187</td>
</tr>
<tr>
<td>$\beta_{24}$</td>
<td>-0.74670</td>
<td>0.19485</td>
<td>0.24022</td>
<td>0.24519</td>
</tr>
</tbody>
</table>
3.3 Alternative Estimates

Applying Beran’s canonical modulation estimator for principal components regression amounts to estimating \( X\beta \) by \( X_f(y) = UFU'\gamma \), where \( F \) is a diagonal matrix whose diagonal consists of a string of 1s followed by a (possibly empty) string of 0s, and \( U \) is obtained through singular value decomposition of \( X \). The optimal number of 1s or principal components is found by minimization of an objective function, which is the estimated quadratic risk of estimating \( X\beta \) by \( X_d(F) \). This criterion is of the form:

\[
\hat{R} = \frac{1}{k} \sum_{i=1}^{k} \left[ \sigma^2 f_i + (z_i^n - \bar{z}^2)(1 - f_i)^2 \right]
\]  

(39)

where \( z = U'y \) is a \( k \)-dimensional vector of which \( z_i \) is its \( i \)th element, \( f_i \) is either a 1 or 0, and \( \sigma^2 \) is the LS unbiased estimator of the variance. Computation of the optimal number of principal components reveals that the empirical risk, Eq. (37), is minimized for the square matrix \( \hat{F} \) with a string of 16 diagonal 1s followed by 9 diagonal 0s, and all off-diagonal elements are 0. The value of the empirical risk corresponding 16 principal components is 178.9882.
For ridge regression, the estimated risk expression \( \hat{R} \) is still of the same form, Eq. (3.9); however, \( F \) is no longer a diagonal matrix of 1s and 0s, but instead,

\[
F = \text{diag}[l_i^2 / (l_i^2 + \kappa)]
\]

where \( \kappa \) is the ridge regression tuning parameter and \( l_i \) is the \( i \)th singular value obtained through decomposition of \( X \). Constrained minimization \( (\kappa \geq 0) \) of the above estimated risk as a function of \( \kappa \) yields the optimal solution of \( \kappa^* = 6.68 \) and a minimized objective of 440.90. The matrix \( \hat{F} \) is then obtained from Eq. (40) where \( \kappa = 6.68 \).

Figure 2, as before, depicts a scatter plot on which DBIT, best principal component, and best ridge regression fits are included. Visual inspection alone makes it difficult to determine which method yields superior fit. Principal components (PC), LS, and the DBIT approach all perhaps seem to outperform ridge regression. Numerically, the estimated risks provide a more accurate means of evaluation. Accordingly, the estimated risks are 545.9197, 545.9045, 178.9882, and 440.9017 where the last two estimates are for PC and ridge regression methods. The DBIT and

![FIGURE 2 Comparative fits for the simulated motorcycle data.](image-url)
LS risks are close indeed, reflecting the nearly identical parameter estimates.

To study further the fit produced by the DBIT method, we refer to Fig. 3, which contains two plots; the first shows the unadjusted residuals against the times $t_i$ while the second shows the unadjusted residual values against the fitted means $\hat{y}_i$. The first plot indicates that the residual values are very close to zero for measurement times up to 18, confirming a very close fit of the regression line. From times of 18 to 50, the dispersion around zero increases as the fit of the regression worsens. The outliers are indicated by a poor fit and large residual value, e.g., greater than 50 in absolute value. The second plot shows the residual values against the fitted values. No apparent systematic pattern can by discerned from mere visual inspection. The plot appears as a blur centered around the zero axis. However, outliers can also be identified by residual values greater than 50 in absolute value, which corresponds to Figs. 1 and 2.

From the preceding analysis it becomes evident that the DBIT method fits the motorcycle data reasonably well under the discussed
criteria. Since the sample size is large, $n = 133$, limiting results suggest that these estimates should be within close approximation to the unobserved location parameters as well as LS estimates, a conclusion evidenced from the regression results. Next, we investigate the performance of the DBIT estimator against LS using a smaller subset of the data. Theory tells us that in this case DBIT and LS regression results should exhibit substantial differences, especially in ill-conditioned problems.

3.4 Reduced Subset of the Data

As an additional basis for comparison between DBIT methods and LS, we investigate the fit produced using a smaller subset of the data. In particular, the procedure discards the first 80 observations of the data, and instead uses the remaining 53 observations in the statistical model, Eq. (3.8) (see Fig. 4). It is important to note that the model contains only 28 free parameters and is highly ill-conditioned with $\kappa(X'X) = 3.260466e + 17$ such that a particular column vector of $X$ is nearly an exact linear combination of the others. The Monte Carlo evidence presented by van Akkeren and Judge (2001) shows that, even for a strongly moderate level of ill-conditioning, e.g., $\kappa(X'X) = 100$, LS estimates have large variances where a one-sample data analysis can lead to unreliable results.

Table 2 compares LS and DBIT estimates computed with $\gamma = 1$ and $j = 5$, and shows that the estimates are strongly divergent. In contrast, the fitted means $\hat{y}$ are roughly similar, but with LS significantly “overfitting” the scatter plot. The estimated variances are 958.8862 and 1062.606 with a lower value for LS as expected. If we consider the information that has been omitted when using only a subset of our sample, it becomes evident that the DBIT method produces estimates significantly closer to the results of Table 1 than does LS. In this aspect, the DBIT method performs better than LS in that the results produced seem reasonable in comparison to the larger data set. Since the true location parameters are unobservable, we rely on asymptotic theory and consistency of the above estimates to conclude that the estimates in Table 1 are more accurate than those produced in Table 2.

3.5 Canadian Earnings Data

The following set of data is from Ullah (1985) and is analyzed by Chu and Marron (1991). Figure 5 shows the scatter plot of 205 data points of log (income) against age where an obvious upward sloping trend followed by a decreasing bend is observable. To form a relationship between earnings
and age, the following polynomial regression model is proposed:

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}^2 + e_i \quad \text{for} \quad 1 \leq i \leq 205 \]  

(41)

The usual assumptions regarding the linear statistical model are valid (Section 1): the unknown errors \( e_i \) are independent with mean 0 and variance \( \sigma^2 \), the design matrix \( X \) is considered nonstochastic, and the unknown location parameters \( \beta \in B \) are finite valued.

### 3.6 Estimation Results

In applying the DBIT estimation procedure to the Canadian earnings data set, the same computational difficulty arises as in the previous example in that the number of unknown variables associated with 205

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DBIT((\gamma = 1))</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0)</td>
<td>3.30360</td>
<td>-2.30566e + 9</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>-5.36008</td>
<td>-1.78241e + 9</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>5.64011</td>
<td>2.45064e + 9</td>
</tr>
<tr>
<td>(\beta_3)</td>
<td>-1.53831</td>
<td>2.64337e + 9</td>
</tr>
<tr>
<td>(\beta_4)</td>
<td>2.34618</td>
<td>2.20861e + 8</td>
</tr>
<tr>
<td>(\beta_5)</td>
<td>-3.41809</td>
<td>-9.25645e + 8</td>
</tr>
<tr>
<td>(\beta_6)</td>
<td>1.18621</td>
<td>-4.58583e + 8</td>
</tr>
<tr>
<td>(\beta_7)</td>
<td>-3.98931</td>
<td>3.79388e + 7</td>
</tr>
<tr>
<td>(\beta_8)</td>
<td>6.98036</td>
<td>8.46492e + 7</td>
</tr>
<tr>
<td>(\beta_9)</td>
<td>-5.03064</td>
<td>1.73084e + 7</td>
</tr>
<tr>
<td>(\beta_{10})</td>
<td>-1.49742</td>
<td>-2.18760e + 6</td>
</tr>
<tr>
<td>(\beta_{11})</td>
<td>2.12458</td>
<td>-8.84952e + 5</td>
</tr>
<tr>
<td>(\beta_{12})</td>
<td>-9.39358</td>
<td>-3.72874e + 4</td>
</tr>
<tr>
<td>(\beta_{13})</td>
<td>-3.21095</td>
<td>-3.97916e + 9</td>
</tr>
<tr>
<td>(\beta_{14})</td>
<td>7.38786</td>
<td>-2.74960e + 9</td>
</tr>
<tr>
<td>(\beta_{15})</td>
<td>-7.52927</td>
<td>8.32280e + 8</td>
</tr>
<tr>
<td>(\beta_{16})</td>
<td>4.86897</td>
<td>1.83482e + 9</td>
</tr>
<tr>
<td>(\beta_{17})</td>
<td>-7.64678</td>
<td>5.63959e + 8</td>
</tr>
<tr>
<td>(\beta_{18})</td>
<td>5.77725</td>
<td>-3.08928e + 8</td>
</tr>
<tr>
<td>(\beta_{19})</td>
<td>-3.03776</td>
<td>-2.38434e + 8</td>
</tr>
<tr>
<td>(\beta_{20})</td>
<td>-0.16285</td>
<td>-2.45760e + 7</td>
</tr>
<tr>
<td>(\beta_{21})</td>
<td>1.07573</td>
<td>1.93329e + 7</td>
</tr>
<tr>
<td>(\beta_{22})</td>
<td>-1.80745</td>
<td>5.39118e + 6</td>
</tr>
<tr>
<td>(\beta_{23})</td>
<td>6.89803</td>
<td>7.20959e + 4</td>
</tr>
<tr>
<td>(\beta_{24})</td>
<td>5.02754</td>
<td>-5.90341e + 4</td>
</tr>
</tbody>
</table>
data points creates a formidable optimization problem. The computational burden is similarly eased through application of ordered quantiles for which \( j = 5 \) is selected. To address the model selection problem of finding the optimal number of polynomial regressors, the Mallows' \( C_p \) (Mallows, 1973) criterion is applied, which penalizes the residual sum of squares by twice the number of parameters times the residual mean square for the initial model (Venables and Ripley, 1994, Chap. 10). The final model is chosen through a sequence of forward or backward steps, which minimizes this criterion. Specifically, we have

\[
C_p = \text{RSS}_k + 2p \hat{s}^2
\]  

for \( p \) parameters of the submodel, and \( \hat{s}^2 \) is the unbiased estimate of the variance in the full model.

Applying backward model selection, LS under the above \( C_p \) criterion derives an optimal model with five polynomial regressors, whereas the DBIT approach produces a final model with four explanatory variables. This result may very well be a consequence of rounding error of using the GAMS optimization package in conjunction with the large values

![FIGURE 4](image-url) Comparative fits for the reduced motorcycle data, \( n: 81\text{--}133 \).
produced in fourth to fifth degree polynomials. Table 3 summarizes the main results. It is observed that DBIT estimates under $\gamma = 0.85$ are very close indeed to LS results. While coefficients derived under the standard scale parameter of $\gamma = 1$ are significantly different, a reasonable fit is produced as can be seen from Fig. 5. The large sample size of the data set leads us to suggest that the DBIT and LS estimates should be fairly similar, a result supported by Table 1.

Figure 5 shows several fits attained under different values of the scale parameter $\gamma$. While each of the three fits seems reasonable, the coefficients corresponding to $\gamma = 0.85$ and $\gamma = 10$ are significantly different. From visual inspection, it appears that the DBIT fitted values under $\gamma = 0.85$ describe the variation of the data best, accounting for the dips and turning points in the data. The dip in fitted values between age of 33 to 50 represents a decrease in income found by Ullah (1985) to be significant and cannot be readily observed from visual inspection. From a labor economics perspective, this observation represents an interesting issue in whether this phenomenon is truly due to some underlying economic structure or merely a coincidental sampling artifact.
3.7 Smoothing Methods

In addition to semiparametric methods such as the polynomial regression model, the data may be fit using nonparametric regression techniques or scatter plot smoothers. The advantage of this approach is that the linear regression model is replaced by the more general equation:

\[ y = \beta_0 + \sum_{j=1}^{k} f_j(x_j) + e \]  

(43)

where \( f \) is a twice-differentiable function that includes the linear regression model. Generally, these methods produce estimates that result from a compromise between goodness of fit and degree of smoothness. A variety of nonparametric regressors exist, but for comparison purposes to the DBIT approach, we will present only the smoothing spline and the kernel regression method.

The cubic smoothing spline, S-Plus command `smooth.spline`, selects the function \( f \) in Eq. (43), which minimizes a weighted functional between goodness of fit and smoothness of the form:

\[ S_l(f) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx \]  

(44)

where \( \lambda \) is a degree of freedom parameter that regulates the amount of smoothness and \( \int [f''(x)]^2 dx \) is the roughness penalty used to quantify local variation. The minimization of \( S_l(f) \) over the class of twice-differential functions defined on the interval \([a, b] = [x_1, x_n] \) has the unique solution, \( \hat{m}_f(x) \), which is a cubic spline (polynomial) due to the particular choice of the roughness penalty. Several methods exist for selecting an optimal parameter, but as before in our case visual inspection is used to

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \gamma = 0.85 )</th>
<th>( \gamma = 1 )</th>
<th>( \gamma = 10 )</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-19.124700</td>
<td>-13.498690</td>
<td>8.126709</td>
<td>-19.177150</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>3.198929</td>
<td>2.592379</td>
<td>0.282846</td>
<td>3.204908</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.114169</td>
<td>-0.090809</td>
<td>-0.002351</td>
<td>-0.114413</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>0.001771</td>
<td>0.001386</td>
<td>0.000055</td>
<td>0.001775</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>-0.000010</td>
<td>-0.000007</td>
<td>0.000001</td>
<td>-0.000010</td>
</tr>
</tbody>
</table>
determine \( \lambda \). Another nonparametric method, the kernel regression estimator, Splus command \textit{ksmooth}, has the specific functional form:

\[
\hat{y}_i = \frac{\sum_{j=1}^{n} y_j K\left(\frac{x_i - x_j}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b}\right)}
\]

Here, as in the case of density estimation, \( K \) is the kernel function for which we shall use the standard normal density \( N(0, 1) \), and \( b \) is the bandwidth parameter that determines the amount of smoothness. The kernel can be any continuous, symmetric, and bounded function that must integrate to one. Choice of optimal bandwidth amounts of finding the optimal balance between bias versus variance of Eq. (45), often called the Nadaraya–Watson estimator (Nadaraya, 1964; Watson, 1964). Again, the optimal value of \( b \) is selected using mere visual inspection.

Figure 6 displays the two smoothing methods compared to the DBIT results attained using optimal polynomial regression under the Mallows’ \( C_p \) criterion. The smoothing spline parameter \( \lambda \) is set equal to 15 with the kernel regression parameter \( b \) equal to 3.5. The plot at first glance indicates that the curves are fairly similar with the characteristics

**FIGURE 6** Scatter plot and smooths for Canadian earnings data.
of the data well reflected by each fit. The DBIT method seems to pick up the initial increase in earnings from age 20 to 30 better than \textit{ksmooth} but not as well as the cubic smoothing spline. The dip in earnings is well observed by each method; however, the DBIT curve seems to weight the final data point in excess, producing a steeper slope than with alternative methods.

4 CONCLUDING REMARKS

Economic data are to a large extent nonexperimentally generated. Since the data, the statistical model, and the corresponding estimation and inference procedure are interdependent, we are restricted to the process of nonexperiential model building. This means there is uncertainty relative to the probability model that best describes the sampling process for the observed data and little basis for specifying the underlying sampling distribution. In addition, since the data are nonexperiential in nature, there may not be enough information contained in the design matrix $X$ and the noisy data $y$ to permit the recovery of the $k$-dimensional $\beta$ vector by traditional LS or maximum likelihood estimation methods. The ill-posed aspect may arise because of nonstationary or other model specification reasons and may cause the number of parameters to exceed the number of data points or the moment conditions to be less than or greater than the number of unknown parameters. In addition, the design matrix $X$ implied by nature or society may be ill-conditioned. Consequently, if traditional estimation procedures such as LS and maximum likelihood methods are used, the solution may be undefined or the estimates may be highly unstable, giving rise to high variance and low precision for the recovered parameters. As a result of this problem, the DBIT estimator and the MOR methods as unbiased estimators of the risk are suggested as alternative methods of estimation.

The two data examples have demonstrated the ability of the DBIT estimator to perform well in comparison with existing methods such as nonparametric smoothing techniques and the MOR approaches. As indicated, this method may be simplified to include quantiles instead of the entire vector of targets when using relatively large sets of data, thereby reducing the computational burden of the constrained optimization problem significantly. In computing the point estimates, the GAMS optimization software package is used to solve the empirical DBIT optimization problem in a matter of seconds for the given sample sizes. The DBIT formulation is shown to be computationally tractable, especially in the quantile formulation. In large samples, DBIT estimates are shown to approach LS results. However, LS tends to “overfit” and
produce unreliable point estimates in contrast to our method, especially when ill-conditioning of the design matrix is a considerable problem as evidenced by the analysis of the reduced Silverman motorcycle data set. One possible implication of the MOR approach is that one may consider the case of applying a biased Stein’s estimator of the risk, a formulation that could lead to desirable sampling properties when the data exhibit collinearity problems.

ACKNOWLEDGMENTS

George Judge and Rudy Beran provided valuable insights and feedback during the course of this research.

REFERENCES


R. Glen Donaldson
University of British Columbia, Vancouver, British Columbia, Canada

Mark J. Kamstra
Federal Reserve Bank of Atlanta, Atlanta, Georgia, U.S.A.

1 INTRODUCTION

A core tenet of financial economics states that, in a market populated by rational investors, the fundamental price of an asset equals the expected discounted present value of its future cashflows. This implies that in a rational and efficient market, and in the absence of price bubbles,* stock price movements are driven by forecasted changes in dividends and discount rates and not by the “irrational exuberance” of traders. Interest in the extent to which stock prices are fundamentally driven has been particularly high of late, with everyone from academics to the Chairman of the U.S. Federal

*The term “bubble” is formally defined below, but for now can be thought of as a situation in which some force other than a “fundamental factor,” such as dividends and discount rates, drives stock prices.
Reserve Board offering opinions. Research into fundamental valuation is, therefore, particularly timely and, at least for the type of computer-intensive research we undertake in this chapter, is made possible by recent increases in computing power available to financial econometricians.

Several different procedures have been developed to estimate fundamental stock prices and to test whether market prices deviate in significant ways from the estimated fundamental price. Many studies have reported what appear to be important deviations from fundamentals, especially around market crashes such as those which occurred in 1929 and 1987. The resulting belief that financial markets may be “excessively volatile,” and may potentially contain “price bubbles” that push market prices away from fundamental valuations, has contributed to the institution of policies such as trading halts in the face of large price moves. Whether such policies help or hurt financial markets depends in important ways on the extent to which market prices are driven by fundamental factors as opposed to “irrational exuberance,” and the answer to this question crucially depends on the accuracy of the fundamental price estimates and tests used in the analysis.

To the best of our knowledge, no one has yet undertaken a thorough and systematic investigation of the accuracy of various fundamental price estimating and testing procedures.* Thus, while a fundamentals-estimation exercise might value a share of stock at $X, we cannot be sure how confident to be in the estimate. For example, the estimate could be biased up or down and/or be very imprecise (i.e., have a large variance). With a small bias and low variance the estimate might be very accurate and therefore of considerable use. Conversely, a significantly biased estimates with very low precision might be wildly inaccurate and thus of little value. Establishing the properties of various fundamentals estimation and testing procedures would, therefore, be of significant benefit to academics, investors, and policy makers who often employ fundamental price estimates in their work. The accuracy of fundamental price estimates can be obtained analytically under only very restrictive and special circumstances

*Bollerslev and Hodrick (1995) provide a survey of much of this literature, and themselves use simulation techniques to investigate tests of market efficiency. Bollerslev and Hodrick (1995) focus much of their simulation study on the constant discount rate case and consider only one ad hoc model for time-varying discount rates. This model, like the constant discount rate case, allows an analytic solution for the market price. We consider formal discount rate processes that admit serial dependence and hence rule out a general analytic solution. Allowing serial dependence in discount rates considerably complicates the analysis, but also considerably enriches the range of possible dynamics for prices and returns.
Estimating and Testing Fundamental Stock Prices

(such as assuming that dividend growth rates and interest rates will never change in the future, which is of course highly unrealistic). To circumvent this analytical intractability, we develop and test in this chapter a simulation-based method for calculating the properties of various fundamental price estimating and testing procedures. We are particularly interested in determining the statistical properties of fundamental price and return estimates commonly used in both industry and academia and in investigating the effects that estimation inaccuracies may have on the variety of volatility tests commonly employed in the literature.

Our strategy in this chapter is as follows: (1) use financial market data to estimate time-series models for dividend growth and discount rates, (2) use these models to simulate dividend growth and discount rate paths for a variety of possible economies that do not contain bubbles, (3) calculate fundamental prices for these bubble-free economies based on the simulated dividend growth and discount rates (which prices we call “market prices” since they are fair-value prices for the simulated market economies), (4) use various fundamental valuation models to estimate fundamental prices for each of the simulated economies, and (5) compare market price versus fundamental price and investigate statistical properties of common tests for excess volatility and bubbles in stock prices using the data we simulated under the (true) null hypothesis that there are no price bubbles. We apply this procedure to S&P 500 stock price data.

Results produced using our computer-aided techniques suggest that, while stock prices are indeed volatile, they are not more volatile than one would reasonably find in an economy driven by fundamentals, thereby suggesting that market prices are not “excessively volatile.” We also find that traditional tests for excess volatility and bubbles overreject a true null of no bubbles in samples of the size traditionally employed in the literature. Indeed, most tests we investigate find overwhelming evidence of bubbles in data series we construct under the conditions that there are no bubbles. In other words, we demonstrate that traditional tests for price bubbles frequently find bubbles in data that do not in fact contain bubbles.

In Section 2, we describe a variety of common fundamental pricing models and tests for excess volatility and bubbles. In Section 3 we present our Monte Carlo procedure for estimating fundamental prices/returns and investigating test performance. We discuss the results of our efforts in Section 4. Section 5 concludes.

2 FUNDAMENTAL PRICING METHODS AND TESTS

Define $P_t^M$ as a stock’s beginning-of-period-$t$ market price, $r_t$ as the rate used to discount payments received during period $t$, and $E_t[\cdot]$ as the
conditional expectation operator, with the conditioning information being
the set of information available to investors at the beginning of period t.
Investor rationality requires that the current market price of a stock, which
will pay a dividend $D_{t+1}$ at the beginning of period $t+1$ and then
immediately sell for the ex-dividend market price $P_{t+1}^M$, satisfies Eq. (1):

$$ P_t^M = E_t \left\{ \frac{P_{t+1}^M + D_{t+1}}{1 + r_t} \right\} $$

We can solve Eq. (1) forward to period $T$ (where $T > t$) and substitute
realized dividends and discount rates in for their expected values to
produce Eq. (2):

$$ P_t^X = \sum_{i=1}^{T-t} \left( \prod_{k=1}^{i-1} \left[ \frac{1}{1 + r_{t+k-1}} \right] \right) D_{t+i} + \left( \prod_{k=1}^{T-t} \left[ \frac{1}{1 + r_{t+k-1}} \right] \right) P_T^M $$

The superscript $X$ on the left-hand-side price indicates that this is the
ex-post rational price; i.e., the price that an investor would have rationally
paid for the stock had she known that the market price of the stock in
period $T$ was going to be $P_T^M$ and that the stock was going to pay the
sequence of dividends that it actually paid between periods $t$ and $T$. The
ex-post rational price is not a fundamental price, nor is it a price that
would ever be observed in the marketplace. However, for reasons
explained below, it is nonetheless useful in analyzing price behavior.

2.1 Fundamental Pricing Models

To calculate the fundamental price of stock, we need to solve Eq. (1)
forward into infinite time under the transversality condition that the
expected present values of the market price $P_{t+i}^M$ falls to zero as $i$ goes
to infinity; i.e., there are no price “bubbles.” This produces the familiar result
that today’s fundamental stock price, $P_t^F$, equals the expected present value
of future dividends, i.e.,

$$ P_t^F = E_t \left\{ \sum_{i=1}^{\infty} \left( \prod_{k=1}^{i} \frac{1}{1 + r_{t+k-1}} \right) D_{t+i} \right\} $$

Note from Eq. (3) that the fundamental price is a function only of
dividends, discount rates, and time, and not of the market price.

Defining the growth rate of dividends during period $t$ as $g_t \equiv (D_{t+1} - D_t)/D_t$ allows the preceding equation to be rewritten as

$$ P_t^F = D_t E_t \left\{ \sum_{i=1}^{\infty} \left( \prod_{k=1}^{i} \frac{1 + g_{t+k-1}}{1 + r_{t+k-1}} \right) \right\} $$
or, defining the discounted dividend growth rate as $y_t \equiv (1 + g_t)/(1 + r_t)$.

$$P^F_t = D_t \mathcal{E}_t \left\{ \sum_{i=1}^{\infty} \prod_{k=1}^{i} y_{i+k-1} \right\}$$

(5)

This can be rewritten as shown in Eq. (6) (note that $y_t$ is not in the time $t$ information set since it depends on $D_{t+1}$ through $g_t$):

$$P^F_t = D_t \mathcal{E}_t \prod_{k=1}^{\infty} y_{i+k-1}$$

(6)

In practice, fundamental valuation calls for forecasting future discounted dividend growth rates (i.e., dividends and discount rates) to solve for the sum product in Eq. (4) [or, equivalently, Eq. (6)]. Since Eq. (4) cannot be solved analytically in general, this usually requires that some restrictive assumptions be made concerning the time-series processes driving dividends and discount rates. The simplest approach, introduced by Gordon (1962), is to assume that discount rates and dividend growth rates will be constant, at $r$ and $g$, respectively, for all future time so that Eq. (4) reduces to Eq. (7) in which the superscript $G$ denotes the Gordon-model fundamental price estimate:

$$P^G_t = D_t \left[ \frac{1 + g}{r - g} \right]$$

(7)

The Gordon model is certainly convenient, but its extremely restrictive assumptions of constant $r$ and $g$ do not tend to produce the most accurate valuations possible. Several attempts have therefore been made to relax Gordon’s original restrictions and yet still retain an analytical solution to Eq. (4). The literature in this area is rather large and includes papers by Malkiel (1963), Fuller and Hsia (1984), Brooks and Helms (1990), Hurley and Johnson (1994, 1998) and Yao (1997), among others. Early literature in this area broke future time into several “chunks,” with dividend growth and discount rates constant within each time chunk, but different between chunks. For example, dividends might be forecasted to grow at a “high rate” for the first 3 years, and then grow at some “normal rate” for the rest of time, so that the sum in Eq. (4) would be broken into two parts, each part solved separately, and then added together (see, e.g., Brooks and Helms, 1990).

Recent work on extending the Gordon model has largely focused on allowing for more complicated time-series behavior of dividends and discount rates, while retaining the ability to solve Eq. (4) analytically. Two particularly good examples, found in Yao (1997), are the additive Markov model (Eq. 1 of Yao, 1997) and the geometric Markov model.
(Eq. 2 of Yao, 1997). These appear below as Eq. (8) and Eq. (9), respectively:

\[ P_t^{ADD} = D_t^{1/r} + [1/r + (1/r)^2] (q^u - q^d) \Delta \]

\[ P_t^{GED} = D_t^{1/r} \left[ 1 + (q^u - q^d) \Delta_{50} \right] \]

\[ \frac{1}{r - (q^u - q^d) \Delta_{50}} \]

in which \( q^u \) is the proportion of the time the dividend increases, \( q^d \) is the proportion of the time the dividend decreases, \( \Delta = \sum_{t=2}^{T} |D_t - D_{t-1}|/(T - 1) \) is the average absolute value of the level change in the dividend payment, and \( \Delta_{50} = \sum_{t=2}^{T} (D_t - D_{t-1})/D_{t-1}/(T - 1) \) is the average absolute value of the percentage rate of change in the dividend payment.

Donaldson and Kamstra (1996, 2000) develop an alternative approach (hereafter referred to as the DK procedure) which does not require that Eq. (4) be solved analytically. The DK procedure instead uses Monte Carlo methods to solve this equation numerically. The DK procedure eases the conditional constancy restrictions on dividend growth rates and discount rates of the Gordon procedure by modeling the discounted dividend growth rate \( y \) in Eq. (6) explicitly as conditionally time varying. The basic idea of the DK procedure is to estimate an econometric model for the time-series behavior of \( y_{t+i} \) (the discounted dividend growth rate) in Eq. (6), which is equivalent to Eq. (4), use this model and randomly drawn innovations to simulate time paths for the possible future evolution of \( y_{t+i} \), then take the present value of the forecasted time paths to find a fundamental price. This is done thousands (even millions) of times, with a different sequence of randomly drawn innovations each time, so as to integrate out the expectation in Eq. (6) and thus produce a numerical (as opposed to analytical) estimate of the fundamental price.
The sequence of steps employed to produce a DK fundamental price estimate are described below and depicted in Exhibit 1.

**Step A**: Use in-sample data to specify and estimate an econometric model for conditionally time-varying $y_t$.

**Step B**: Use the estimated model from Step A and data up to period $t$ to simulate, out-of-sample, possible realizations of $y_{t+i}$: $i = 1, \ldots, I$, where $I$ is chosen to be very large (infinity, were it practical),* for a cross-section of $J$ different possible economies, so that we follow the evolution of $y$ across a panel of $J$ simulated economies over $I$ periods of time. We accomplish this by using the estimated model from Step A to make a conditional mean forecast of $y$, denoted $\hat{y}_t$, then simulate a population of $J$ possible shocks around the mean and add these shocks to $\hat{y}_t$ to produce $y_{t+i}^j : j = 1, \ldots, J$ for the cross-section of $J$ economies at time $t$. We then repeat this procedure through time for each of the $J$ economies, conditioning on $y_{t+i}^j$, to form $\hat{y}_{t+i}^j : i = 1, \ldots, I$. We therefore produce an out-of-sample panel of values of $y$ for $J$ economies stretching out $I$ periods, all of which have been simulated based on the in-sample data.

**Step C**: Calculate the fundamental DK price based on this panel of simulated $y$ using Eq. (6) as follows:

$$P_{DK}^t = D_t \sum_{j=1}^{J} \left( \frac{\sum_{i=1}^{I} \prod_{k=1}^{I} y_{t+i}^j}{J} \right)$$

Note that the preceding equation can be rearranged by bringing the dividend level $D_t$ inside the parentheses so that

$$P_{DK}^t = \sum_{j=1}^{J} \left( D_t \frac{\sum_{i=1}^{I} \prod_{k=1}^{I} y_{t+i}^j}{J} \right)$$

It is interesting to note here that the parenthesized term in the preceding equation—i.e., $D_t \sum_{i=1}^{I} \prod_{k=1}^{I} y_{t+i}^j$ is for economy $j$ the ex-post rational price of the stock from Eq. (2), with the terminal period pushed out to infinity as $I$ goes to infinity. In other words, if an investor stood at the end of time and looked back over history to see that the stock had paid a stream of dividends that had grown at rate $g_{t+i}$ and which were discounted

*Note that as $I$ increases, the product $\prod_{i=1}^{I} y_{t+i}$ converges to zero since $y$ is less than unity in steady state. We found that, in practice, $I = 400$ was easily sufficient to have our simulations converge to the point where increasing $I$ had no impact on the fundamental price. In other words, a dividend received 400 years in the future has essentially no impact on today’s price and thus 400 years in the future is equivalent to infinity for practical purposes in our simulations.
at rate $r^j_{t+i}$ such that the discounted dividend growth rate had been $y^j_{t+i}$ in this $j$th economy, the investor would rationally feel that she should have been willing to pay $P^j_{t+i} = D_t \sum_{i=1}^{\infty} \prod_{k=1}^{i} y^j_{t+k}$ to have purchased the stock back in time $t$ where $P^j_{t+i}$ is commonly referred to by Shiller (1981) and others as the “ex-post rational price” [although when applied to data it is truncated at the end of the data period, as shown in Eq. (2)].

The representation of a single fan of the DK simulation as an ex-post rational price is interesting because it highlights the relationship between the ex-post rational price, the Gordon price, and the DK technique, and therefore provides some insight into the likely properties of the price estimates produced by various methods. For example, it is clearly seen that all Gordon-based methods are restricted versions of the DK technique. The basic Gordon model, for example, imposes conditionally constant dividend growth rates and discount rates such that each of the $j$ economies in the simulation are identical and based on a constant $y$. If these restrictions are invalid and there is good reason to believe that they are indeed invalid—then the Gordon estimator will be biased and inefficient relative to DK.

The ex-post rational price calculation performed by Shiller (1981) and others is also problematic because they estimate a fundamental price based only on one realization of dividends and discount rates: the realization actually observed in the true market. In other words, Shiller’s method stands at the end of time and asks what an investor with perfect foresight would have paid back at date $t$ for a share of stock, had she known the dividends and discount rates that were to occur in the future.* Conversely, a real-life investor looks into an uncertain future when making purchasing decisions and therefore considers a universe of possible economies that might unfold when valuing a share of stock—this is the DK method. The ex-post rational price, based on only one time path, will, therefore, provide an imprecise picture of the thousands of possible economies, and their associated dividend time paths, that were considered when forming the market price at which real-life traders buy and sell stock. The extent to which this imprecision might affect our view on important asset-pricing questions, such as whether market prices are excessively volatile or not, is studied below.

*Of course, if the investor had perfect foresight, and therefore faced no uncertainty, she would not require an equity risk premium so the discount rate she would use would be lower than the rate $r$ used in the ex-post price calculation. This subtlety is typically ignored in the literature that uses ex-post price calculations. For consistency we will also follow the literature here.
2.2 Some Tests Using Fundamental Prices

One of the key features of analytically based fundamental pricing procedures is that the fundamental prices they produce behave differently than observed market prices in important ways. For example, the time path of such fundamental prices is typically much smoother than that of observed market prices, which has led many researchers to conclude that market prices are excessively volatile. Indeed, there is an entire literature devoted to the study of excess volatility in financial markets [reviewed by Camerer (1989) and Cochrane (1992)]. Conversely, Donaldson and Kamstra (1996, 2000) find that the DK procedure produces fundamental stock prices that behave substantially the same as market prices in terms of return volatility. Of course, the evidence presented in all of these papers is based on the comparison of one estimate of the fundamental price (the estimate produced by the model in question) with one realization of the market price (the price series we see in the true market data). Given that the results come from only one price series, there is no way to be truly sure how accurate (biased, precise, etc.) the fundamentals estimate might be in general. The use of computer-aided financial econometric techniques in an investigation of the accuracy of various fundamentals-estimation procedures is, therefore, one important objective of this chapter.

Another objective of this chapter is to investigate the properties of various tests applied to fundamental and actual stock price data, in particular tests for excess volatility and price bubbles. Tests we investigate in this chapter include the following standard tests for price bubbles.

First, Camerer (1989) observes that market prices would boom upward and crash downward with bubbles expanding and collapsing, which would lead market prices to be excessively volatile relative to fundamental prices if the market price contained bubbles. This suggests a test with a no-bubbles null hypothesis that the variance of the percentage rate of change in market prices and the percentage rate of change in fundamental prices are equal, and an alternative hypothesis that the market price contains bubbles and thus that the variance in the percentage rate of change in market prices is greater than the percentage rate of change in fundamental prices.

Second, a direct test of bubbles asks whether the market and fundamental prices are cointegrated (e.g., Campbell and Shiller, 1987). Bubbles in the market price will lead to a nonstationary difference between the market price and the fundamental price estimate; a standard unit root test on this difference therefore tests the null of cointegration. See Dickey and Fuller (1981) and Said and Dickey (1984) for a description of unit root tests.
Third, Mankiw et al. (1985) (MRS) develop two tests for bubbles. They exploit a decomposition of the difference between the ex-post rational price, $P_X^t$, [calculated by inserting realized dividends and discount rates into the present value equation as in Shiller (1981), as seen in Eq. (2)], and the fundamental price estimate $P_F^t$, relative to the observed market price $P_M^t$:

$$
\frac{P_X^t}{C_0} - \frac{P_F^t}{C_0/C_1} = \frac{P_X^t}{C_0} - \frac{P_M^t}{C_0/C_1} + \frac{P_M^t}{C_0} - \frac{P_F^t}{C_0/C_1}
$$

Since the total volatility of the two sides of the preceding equation are equal by definition, the volatility of each individual component on the right-hand side should be less than or equal to the volatility of the left-hand side. The test that compares the volatility of the first term on the right-hand side with the volatility of the left-hand side will be labeled the MRS1 test, while the test which compares the volatility of the second term on the right-hand side with the volatility of the left-hand side will be labeled the MRS2 test. In both cases the null hypothesis of no bubbles (i.e., no excess volatility) is that the term on the left-hand side is more volatile than the right-hand side term to which it is being compared. In considering these tests it is worth asking whether any of them have proper size, or whether they over-reject as some researchers argue. Note that these tests involve a joint null hypothesis that the market price does not contain a bubble and that the fundamental price estimate, produced by some fundamental pricing model, shares similar properties with the market’s true unobserved fundamental price. A finding of systematic over-rejection across simulated market economies (where the null of no bubbles is imposed) would therefore indicate that the model used to produce fundamental price estimates is mis-specified.

In this chapter we employ Monte Carlo procedures to investigate the performance of various fundamentals-estimation methodologies and of simulated market prices, and also of various procedures commonly used to test for excess volatility and bubbles in stock prices. Although our particular application of Monte Carlo methods is new, there is already a significant literature that uses Monte Carlo methods in asset-pricing applications. Indeed, there is body of work that (directly or indirectly) simulates stock prices and dividends, under various assumptions, to investigate price and dividend behavior (e.g., Scott, 1985; Kleidon, 1986; West, 1988a,b; Campbell, 1991; Mankiw et al., 1991; Hodrick, 1992; Timmermann, 1993, 1995; Campbell and Shiller, 1998). However, these studies typically impose restrictions on the dividend and discount rate processes so as to obtain fundamental prices from some variant of the Gordon (1962) model discussed above and/or some log-linear approximating framework.
Rather than impose approximations to solve Eq. (4) analytically, we will instead simulate the dividend growth and discount rate processes directly, and evaluate the expectation through Monte Carlo integration techniques. This approach is computationally burdensome since it requires us to perform a Monte Carlo simulation of a Monte Carlo simulation, but it is the only way to evaluate Eq. (4) without approximation error.* We also take care to calibrate our models to the time-series properties of the data. Dividend growth, for instance, is strongly autocorrelated in the S&P500 stock market data, in contrast to the assumption of a log random walk for dividends often imposed in this literature.

3 SIMULATING FUNDAMENTAL STOCK PRICES

Consider again the pricing relationship of Eq. (4), rewritten below for convenience, in which $P_t$ is the price (for notational convenience we drop the superscript “F”), $D_t$ is the dividend, $g$ is the dividend growth rate, and $r$ is the discount rate:

$$P_t = D_t E_t \left\{ \sum_{i=1}^{\infty} \left( \prod_{k=1}^{i} \frac{1 + g_{t+k-1}}{1 + r_{t+k-1}} \right) \right\}$$

We now execute our strategy of (1) calibrating models to market dividends and interest rates, (2) simulating bubble-free economies using these calibrated models, (3) calculating fundamental prices for these bubble-free economies, (4) forming fundamental price estimates, including the Gordon models and the DK model, and (5) evaluating these fundamental price estimates and tests for bubbles.

3.1 Dividends and Discount Rates

The first step is to estimate time-series models for dividend growth and interest rates so that the Monte Carlo simulations generate dividends and discount rates that match real-world dividends and discount rates. This will allow us to generate nonbubble prices and returns and compare them to real-world prices and returns [e.g., a finding that the real-world prices and returns look significantly different from the simulated (nonbubble) prices].

*There is still Monte Carlo simulation error, but that is random, unlike most types of approximation error, and it can also be measured explicitly. This simulation error is analogous to simple cases such as the simulation error associated with Monte Carlo experiments on the size of a test statistic.
prices and returns would provide evidence of non-fundamental movements in real-world prices and returns; possibly bubbles]. This will also allow us to evaluate conventional fundamental price estimation methods, including variants of the Gordon growth model, and conventional tests for bubbles, such as the MRS tests, to explore the properties of the estimators and test statistics when there are no bubbles in the price.

Our dividend process is calibrated to the S&P 500 stock index annual dividend data, 1952–1998, collected as described in Shiller (1989). The discount rate is defined to be the risk-free interest rate plus a constant equity premium, where the risk-free rate is the interest rate on a 1-year U.S. T-bill as constructed by the U.S. Federal reserve, 1952–1998.* A constant equity premium of 5.77% is added to the risk-free interest rate to produce a discount rate consistent with the stock price data.†

S&P 500 dividends grew at an average rate of 5.5% per year, over 1952–1998, with a standard deviation of 3.7%. As dividend growth rates have a minimum value of −100% and no theoretical maximum, a natural choice for their distribution is the log normal in the S&P 500 data. The logarithm of 1 plus the dividend growth rate has mean 0.0531 and standard deviation 0.035 over 1952–1998. We estimated simple autoregressive moving average (ARMA) time-series models for the logarithm of 1 plus the dividend growth rate and found the best model by the Bayesian information criterion to be a moving average model of order 1 (MA(1)) with the MA(1) coefficient equal to 0.60. Standard tests for normality of this error term do not reject the null of normality,‡ and standard tests for autocorrelation and autoregressive conditional heteroskedasticity (ARCH) fail to reject the null of homoskedasticity and no serial correlation.§

*We choose this measure because, as Cochrane (1992) notes, “there is a long tradition in the volatility test and investment or capital-budgeting literature that measures time-varying discount rates from interest rates plus risk premiums that are constant over time.”

†Finance theory requires that $E[(1 + R)/(1 + i + \pi)] = 1$, where $R$ is the log of one plus the stock return, $i$ is the risk-free rate, and $\pi$ is the equity premium (i.e., $1/[1 + i + \pi]$ is the “pricing kernel”). See, e.g., Campbell et al. (1997). Selection of the equity premium that produces the requisite pricing kernel was accomplished with a grid search, and this leads to a risk premium of 5.77% in the annual S&P 500 data over 1952–1998.


§See Engle (1982) for the seminal treatment of ARCH effects.

These are based on autocorrelations of the residuals and residuals squared.
As for interest rates, since economic theory admits a wide range of possible interest-rate processes there are a variety of models possible, from constant to autoregressive and highly nonlinear heteroskedastic forms. The autoregressive model of order 1 (AR(1)) of the logarithm of interest rates, as described in Hull (1993, p. 408), will be used here as it fits our data well and restricts nominal rates to be positive. Standard specification tests for normality, autocorrelation, and ARCH on the error term from an AR(1) model of the logarithm of interest rates do not reject the null of no misspecification. The 1-year T-bill rates have mean 0.059 and standard deviation 0.03 over 1952–1998. The AR(1) coefficient estimate in the regression of log interest rates on lagged log interest rates equals 0.83. Finally, the error terms from the MA(1) model of log dividend growth rates and log interest rates are correlated, with a correlation coefficient of 0.21.

Properties of fundamental prices and returns produced by Eq. (4) hinge delicately on the modeling of the dynamics of the dividend growth and interest-rate processes. For instance, fundamental prices will equal a constant times the dividend level and fundamental returns will be very smooth over time if dividend growth and interest rates are equal to constants plus independent innovations. However, modeling these data series to capture the serial dependence of dividend growth rates and interest rates observed in the data, as we have done, will typically lead to time-varying price-dividend ratios and variable returns of the sort we see in the S&P 500 stock market data.

### 3.2 Monte Carlo Experiment

We now detail the Monte Carlo experiment by which the price, $P$, is arrived at given conditioning information on the dividend level, $D$, dividend growth rate, $g$, and interest rate, $r$. That is, we detail for the $\ell$th economy ($\ell = 1, \ldots, E$) the formation of the price $P_\ell^t$ given $D_\ell^t$, $g_\ell^t$, and $r_\ell^t$. $P_\ell^t$ is the market price that Eq. (4) states would obtain in economy $\ell$ if the stock market in economy $\ell$ is rational, efficient, and bubble-free. In terms of timing and information, recall that $P_\ell^t$ is the stock’s beginning-of-period-$t$ market price based on fundamental factors, $r_\ell^t$ is the rate used to discount payments received during period $t$ and is known at the beginning of period $t$, $D_\ell^t$ is paid at the beginning of period $t$, $g_\ell^t \equiv (D_{\ell, t+1}^t - D_\ell^t)/D_\ell^t$ and is not known at the beginning of period $t$ since it depends on $D_{\ell, t+1}^t$, and $E_{\ell} \{ \cdot \}$ is the conditional expectation operator, with the conditioning information being the set of information available to investors at the beginning of period $t$. Finally, recall from Eq. (4) that investor rationality requires...
Based on the preceding equation, we generate prices by generating a multitude of possible streams of dividends and discount rates, present-value discounting of the dividends with the interest rates, and averaging the results, i.e., by conducting a Monte Carlo integration. Hence, we produce $P_{e1}^e, e = 1, \ldots, E$ utilizing only dividend growth rates and discount rates. The exact procedure is described below and summarized in Exhibit 2.

### 3.2.1 Simulating Stock Prices for a Rational, Efficient, Bubble-Free Market

**Step 1:** When forming $P_{e1}^e$, the most recent fundamental information available to a market trader would be $g_{e-1}^e, D_{e1}^e$, and $r_{e1}^e$. The quantities $g_{e-1}^e, D_{e1}^e$ and $r_{e1}^e$ must, therefore, be generated directly in our Monte Carlo experiment, whereas $P_{e1}^e$ must be calculated based on these $g, r$, and $D$ since this is how prices would be determined by a rational market participant in a bubble-free economy. In the steps below the risk free T-bill rate is indicated as $r$ and the discount rate (i.e., the risk-free interest rate plus the risk premium of 5.77%) as $r_e$. The objective of Steps 1(a)–(c) is to produce dividend growth rates and interest rates that replicate the real-world dividend growth and discount rate data. That is, the simulated dividend growth rates and interest rates must have the same mean, variance, correlation structure, and autocorrelation structure as the real-world dividend growth rates and interest rates.

**Step 1(a):** Note that since the logarithm of one plus the dividend growth rate is modeled as a MA(1) process, $\log(1 + g_{e1}^e)$ is a function of only innovations, labeled $\epsilon_g$. Note also that since the logarithm of the interest rate is modeled as an AR(1) process, $\log(r_{e1}^e)$ is a function of
log(\(r_{t-1}\)) and an innovation labeled \(\varepsilon_{t}^r\). Set the initial dividend, \(D_{1}^e\), equal to the S&P 500's dividend value for 1951 (observed at the end of 1951), and the lagged innovation of the logarithm of the dividend growth rates \(\varepsilon_{t-0}^g\) to 0. To match the real-world interest rate data, set \(\log(r_{0}^e) = -3.05\) (the mean value of log interest rates required to produce interest rates matching the mean and variance of observed T-bill rates). Set the standard deviation of the innovation to the log interest rate process to 0.242, and the standard deviation of the innovation of the log dividend growth rate process to 0.0305. Then, generate two independent standard normal random numbers, \(\varepsilon_{t-1}^r\) and \(\varepsilon_{t-1}^g\), and form two correlated random variables, \(\varepsilon_{t}^r = 0.242(0.21\varepsilon_{t-1}^r + (1 - 0.21)^{0.5}\varepsilon_{t-1}^g)\) and \(\varepsilon_{t}^g = 0.0305\varepsilon_{t-1}^g\). These are the simulated innovations to the interest rate and dividend growth rate processes, formed to have standard deviations of 0.242 and 0.0305, respectively, to match the data, and to be correlated with correlation coefficient 0.21 as we find in the S&P 500 and T-bill data. Next, form \(\log(1 + g_{t}^e) = 0.0531 + 0.60\varepsilon_{t-0}^g + \varepsilon_{t-1}^g\) and \(\log(r_{t}^e) = -0.18 + 0.94\log(r_{0}^e) + \varepsilon_{t-1}^r\). Also, form \(D_{t}^e = \frac{D_{1}^e(1 + g_{t}^e)}{C_{0}}\).

\[ \text{Step 1(b):} \text{ Produce two correlated normal random variables, } \varepsilon_{t}^r \text{ and } \varepsilon_{t}^g \text{ as in Step 1(a), and conditioning on } \varepsilon_{t-1}^g \text{ and } \log(r_{t-1}^e) \text{ from Step 1(a) produce } \log(1 + g_{t}^e) = 0.0531 + 0.60\varepsilon_{t-0}^g + \varepsilon_{t-1}^g, \log(r_{t}^e) = -0.18 + 0.94\log(r_{0}^e) + \varepsilon_{t-1}^r \text{ and } D_{t}^e = \frac{D_{1}^e(1 + g_{t}^e)}{C_{0}}. \]

\[ \text{Step 1(c):} \text{ Repeat Step 1(b) to form } \log(1 + g_{t}^e), \log(r_{t}^e) \text{ and } D_{t}^e \text{ for } t = 3, 4, 5, \ldots, T \text{ and for each economy } e = 1, 2, 3, \ldots, E, \text{ then calculate the dividend growth rate } g_{t}^e \text{ and the discount rate } r_{t}^e = r_{t}^e + 0.0577. \]

\[ \text{Step 2:} \text{ For each time period } t = 1, 2, 3, \ldots, T \text{ and economy } e = 1, 2, 3, \ldots, E \text{ we must calculate present value prices, } P_{t}^e. \text{ In order to do this we must solve for the expectation of the infinite sum of discounted future dividends conditional on time } t-1 \text{ information for economy } e. \text{ That is, we must produce a cross-section of dividends and interest rates that might be observed in periods } t, t+1, t+2, \ldots \text{ given what is known at period } t-1 \text{ and use these to solve the expectation of Eq. (4). The counter } j \text{ below indexes the cross-section of future economies that could possibly evolve from the current state of the economy.} \]

---

*Notice that the AR(1) parameter for the log interest rate process is estimated to be 0.83, but we have set it to 0.94 in the simulations. It is well known that the coefficient estimate in an AR(1) OLS regression is biased downwards; see, for instance, Kennedy (1992, p. 147). Monte Carlo experiments were employed to determine the appropriate correction for our data, as in Orcutt and Winokur (1969), and this led to the setting of 0.94. The intercept term had to be adjusted as well to reflect this new setting.
Step 2(a): Set \( e_{g,t-1}^{j,e} = e_{g,t-1}^e \) and \( \log(r_{j-1}^{e}) = \log(r_{j-1}^e) \) for \( j = 1, 2, 3, \ldots, J \). Generate two independent standard normal random numbers, \( e_{g,t}^{j,e} \) and \( e_{g,t}^{e} \) and form two correlated random variables \( e_{g,t}^{j,e} = 0.242(0.21 e_{g,t-1}^{j,e} + (1 - 0.21^2)^{0.5} e_{g,t}^e) \) and \( e_{g,t}^{e} = 0.0305 e_{g,t}^{j,e} \) for \( j = 1, 2, 3, \ldots, J \). These are the simulated innovations to the interest rate and dividend growth rate processes, respectively. Form \( \log(1 + g_{i,t}^{j,e}) = 0.0531 + 0.60 e_{g,t}^{j,e} + e_{g,t}^{e} \) and \( \log(r_{i}^{j,e}) = -0.18 + 0.94 \log(r_{i-1}^{j,e}) + e_{g,t}^{j,e} \).

Step 2(b): Produce two correlated normal random variables \( e_{r,t+1}^{j,e} \) and \( e_{r,t+1}^{e} \) in Step 2(a), and conditioning on \( e_{g,t}^{j,e} \) and \( \log(r_{j-1}^{e}) \) from Step 2(a) produce \( \log(1 + g_{i,t+1}^{j,e}) = 0.0531 + 0.60 e_{g,t}^{j,e} + e_{r,t+1}^{e} \), and \( \log(r_{i,t+1}^{j,e}) = -0.18 + 0.94 \log(r_{i,t}^{j,e}) + e_{r,t+1}^{j,e} \), for \( j = 1, 2, 3, \ldots, J \).

Step 2(c): Repeat Step 2(b) to form \( \log(1 + g_{i,t+1}^{j,e}) \) and \( \log(r_{i,t+1}^{j,e}) \) for \( i = 2, 3, 4, \ldots, I \), \( j = 1, 2, 3, \ldots, J \), and economies \( e = 1, 2, 3, \ldots, E \). Solve for the dividend growth rate \( g_{i,t+1}^{j,e} \), the dividends \( D_{i,t}^{j,e} \), and the discount rate \( r_{i,t+1}^{j,e} = r_{i,t}^{j,e} + 0.0577 \) for \( i = 0, 1, 2, \ldots, I \).

Step 2(d): The present discounted value of each of the individual, \( J \) streams of dividends is now taken in accordance with Eq. (4). Note that for each of the \( J \) streams the present value price so calculated for that stream is the ex-post rational price for the \( j \)th economy; the price a rational investor would pay for the stock if she knew for certain that the \( j \)th economy would obtain. We can, therefore, call the \( j \)th present value \( P_{i}^{X,j,e} \), where the superscript \( X \) denotes ex-post rational.

In considering these prices, note that according to Eq. (4) the stream of discount and dividend growth rates should be infinitely long, while in our simulations we extend the stream only a finite number of periods, \( I \). Since the ratio of gross dividend growth rates to gross discount rates, i.e., the \( g \)'s in Eq. (6) are less than one in the steady state, the individual product elements in the infinite sum in Eq. (6) [equivalently, Eq. (4)] eventually converge to zero as \( I \) increases. Indeed, this convergence to zero is exactly what is required for the absence of price bubbles! We therefore set \( I \) large enough in our simulations so that the truncation does not materially effect our results. As mentioned previously, we found that setting \( I = 400 \) (years) accomplishes this conservatively. That is, the discounted value today of a dividend payment received 400 years in the future is essentially zero. Also note that the steps above are required to produce \( P_{i}^{X}, D_{i}^{j}, g_{i,t}^{e-1}, \) and \( r_{i,t}^{e} \) for \( e = 1, \ldots, E \); the intermediate terms

*For our random number generation we made use of a variance reduction technique, stratified sampling. This technique has us drawing pseudorandom numbers, ensuring that \( q\% \) of these draws come from the \( q\% \) percentile, so that our sampling does not weight any grouping of random draws too heavily.
superscripted with a $j$ are required only to perform the numerical integration that yields $P_i^j$.

**Step 2(e):** Perform Steps 1(a)–(c) and 2(a)–(d) for $t = 1, \ldots, T$, rolling out $E$ independent economies for $T$ periods. The length of the time series $T$ is chosen to be 47 to imitate the 47 years of annual data we have available from the S&P 500 from 1952–1998.* This produces $D_i^e$, $r_i^e$, and $P_i^e$, for $t = 1, 2, 3, \ldots, T$ and $e = 1, 2, 3, \ldots, E$. We can also construct the market returns for these economies:

$$R_i^e = \frac{(P_i^{e+1} + D_i^{e+1} - P_i^e)}{P_i^e}$$

and the equity premium, $\pi^e$, that agents in the $e$th economy would observe. The equity premium satisfies the pricing-kernel condition:

$$E \left\{ \frac{1 + R_i^e}{1 + r_i^e + \pi^e} \right\} = 1$$

Exhibit 3 summarizes this procedure by which we calculate market prices for the $E$ economies in Step 2(e).

---

*To avoid initial conditions contaminating the simulations, the dividend growth rates and interest rates are simulated for over 100 periods before dividends and prices are calculated.

†As with determining the equity premium for the S&P 500 data, this requires a nonlinear estimation problem to be solved and was accomplished with a grid search. The equity premium is restricted to be positive and thus is set to 0 if $\pi^e < 0$. 

---

**EXHIBIT 3** Diagram of the $e$th economy of $E$ economies and calculation of the $e$th economy market price at time $t+k$ [Step 2(e)].
3.2.2 Calculating Fundamental Prices

Now that we have calculated market prices for each of the $E$ economies, we next move to calculate estimated fundamental prices for the same economies based on various fundamentals-estimation procedures, such as the Gordon and DK models.

**Step 3:** It is useful for future reference first to calculate ex-post rational prices for these $E$ economies; these are the prices that obtain by substituting realized interest rates and dividends for their expected values in Eq. (2). To do this define $P^{\text{ex}}_{t,e}$ as the ex-post rational price for economy $e$ at time $t$ and treat $P^e_T$ as the truncation point price, i.e., the last price observation in the sample.* We will use the realized dividends $D^e_t$, interest rates $r^e_t$ and equity premium $\pi^e$ to perform the calculation for each economy $e = 1, 2, 3, \ldots, E$. This calculation of ex-post rational prices for each of the $E$ economies and $t = 1, \ldots, T$ time periods produces $E$ time series of ex-post rational prices produced by our simulations.

**Step 4:** We must also form Gordon-model prices for each of these $E$ economies. That is, for each of the $e = 1, 2, 3, \ldots, E$ simulated economies for which market prices were calculated in the previous section, we now calculate a fundamental price using the Gordon model.

**Step 4(a):** Set $\bar{g}^e = \sum_{t=1}^T g^e_t / T$ and $\bar{r}^e = \sum_{t=1}^T r^e_t / T + \pi^e$, and form the Gordon price:

$$P^{\text{ex}}_{t,e} = D^e_t \left[ \frac{1 + \bar{g}^e}{\bar{r}^e - \bar{g}^e} \right]$$

**Step 4(b):** In addition to the classic Gordon model, there have been a variety of extensions to make the Gordon model more realistic by allowing dividend growth rates to vary over time. We implement the trinomial dividend models of Yao (1997), the additive Markov model, and the geometric Markov model specified in Eqs. (8) and (9).

The additive model has us estimating for each $e = 1, 2, 3, \ldots, E$ economy the average absolute value of the change in the dividend payment. $\Delta^e = \sum_{t=2}^T |D^e_t - D^e_{t-1}| / (T - 1)$, the proportion of the time the dividend increases, $q^{\text{inc}}$, and the proportion of the time the dividend decreases, $q^{\text{dec}}$, in order to form the price estimate:

*In the literature (e.g., Shiller, 1981), when calculating ex-post rational prices, a truncation point is chosen to be either fixed, at say the last observed market price, or to be a moving point 20 or more years ahead of the date for which we are calculating the price. With a fixed date, the ex-post converges to the market price at the terminal date.
The geometric model has us estimating for each economy \( e = 1, 2, 3, \ldots, E \) the average absolute value of the percentage change in the dividend payment, \( \Delta^{e,\%} = \sum_{t=2}^{T} |(D_t^e - D_{t-1}^e)/D_{t-1}^e|/(T-1) \), to form the price estimate:

\[
P_{t^{\text{GEO}},e} = D_{t-1}^e \left[ \frac{1 + (q^{e,u} - q^{e,d})\Delta^{e,\%}}{r^e - (q^{e,u} - q^{e,d})\Delta^{e,\%}} \right]
\]

**Step 5:** Finally, we also consider the Donaldson and Kamstra (1996) procedure, which focuses on the ratio of dividend growth to discount rates, \( \log(y_t) \equiv \log((1 + g_t)/(1 + r_t + \pi)) \), rather than on dividend growth and interest rates separately. The DK procedure calls for evolving many possible streams of \( y \)’s into the future with a Monte Carlo simulation and then taking the present value as in Eq. (10). Donaldson and Kamstra (1996) argue that \( y_t \) is better behaved than \( g_t \) or discount rates alone and that it makes more sense to forecast \( y_t \) since this is the object of investor interest in Eq. (6). In other words, investors care about the ratio of gross dividend growth to the gross discount rate, not about each variable individually, so it makes more sense to forecast the ratio \( y \). The DK procedure is described below and summarized in Exhibit 4.

**Step 5(a):** To implement the DK procedure, we start by estimating a model for \( \log(y) \) that captures its dynamic evolution over time. For each of the \( e = 1, 2, 3, \ldots, E \) simulated economies for which market prices were calculated in the previous section, we now use the Bayesian information criterion to select the “best” model from the set of ARMA \((p,q)\) models, \((p,q) = (1,0), (1,1), \text{ and } (2,0) \) for \( \log(y_t) \). We find that \((p,q) = (1,0)\) is

---

**EXHIBIT 4** Diagram of DK Monte Carlo integration of the \( e \)th economy.
usually, but not always, chosen. For each of the $E$ economies we then use the estimated model to make conditional mean forecasts $\log(y_j^t)$, $t = 1, \ldots, T$, conditional on only data observed before period $t$ for that economy. We also estimate the sample standard deviation ($\hat{\sigma}^e$) of the residual of the model within each economy. With $E = 1000$ economies we would do this step 1000 times, once for each of the $e = 1, \ldots, E$ economies.

**Step 5(b):** Now simulate discounted dividend growth rates for each of the $e = 1, 2, 3, \ldots, E$ simulated economies. That is, produce $\log(y^e)$ that might be observed in period $t$ in economy $e$ given what is known at period $t-1$ in economy $e$. To do this for a given period $t$ and economy $e$, simulate a population of $J$ independent possible shocks (draws from a normal distribution,* mean 0, and standard deviation equal to $\hat{\sigma}^e$), which we label $e_i^{j,e}$, $j = 1, \ldots, J$, and add these shocks separately to the conditional mean forecast $\log(y_j^t)$ from Step 5(a) so as to produce $\log(y_j^{t,e}) = \log(y_j^t) + e_i^{j,e}$, $j = 1, \ldots, J$. This is a simulated cross-section of $J$ possible realizations of $\log(y_j^t)$ considered at time $t-1$ for economy $e$, i.e., different paths that economy $e$ may take next period. Note here that we are performing a Monte Carlo simulation on each of the $e$ economies that were themselves generated by a Monte Carlo simulation, so that if we generate $E = 1000$ economies in Steps 1–2, and $J = 1000$ economies in Step 5, we are in total generating one million economies for the DK simulation (which is precisely what we have done for the investigations in this chapter).

**Step 5(c):** Next, for each economy $e$, use the estimated model from Step 5(a) to make the conditional mean forecast $\log(y_j^{t,e})$, conditional on only the $j$th realization for period $t$, $\log(y_j^{t,e})$, and $e_i^{j,e}$, and the data known at period $t-1$, and simulate a population of $J$ independent shocks $e_i^{j,e}$, $j = 1, \ldots, J$ as in Step 5(b) to form $\log(y_j^{t+1,e})$.

**Step 5(d):** Repeat Step 5(c) to form $\log(y_j^{t+1,e})$, $\log(y_j^{t+2,e})$, $\ldots$, $\log(y_j^{t+I,e})$ for each of the $J$ economies, where $I$ is the number of periods into the future the simulation is run ($I = 400$ in our simulations, as explained above). Then form the simulated ex-post rational price, $P_t^{j,b,e}$, where

$$P_t^{j,b,e} = D_t(y_{j_1}^{t,e} + y_{j_2}^{t,b,e} y_{j_3}^{t+1,e} + y_{j_4}^{t,b,e} y_{j_5}^{t+2,e} + \cdots)$$. $j = 1, \ldots, J$

*The regression error from time-series models of $\log(y)$, formed with the S&P 500 and T-bill data, are normally distributed, leading us to this choice of distribution for the shocks here. Limited experiments with the Monte Carlo $\log(y)$ series indicates that the regression error invariably appears to be normally distributed. As $y$ is a ratio of log normal random variables, this is not surprising.
corresponding to the $J$ possible economies based on the previously simulated economy $e$.

**Step 5(e):** Calculate the DK fundamental price for each time period $t = 1, \ldots, T$ and each economy $e = 1, 2, 3, \ldots, E$:

$$P_{t}^{DK, e} = \sum_{j=1}^{J} P_{t}^{i, j, e} / J$$

The DK procedure outlined above is represented diagrammatically in Exhibit 4. For the experiments in this chapter we set $E = 1000$, $J = 1000$, and $I = 400$.*

### 3.2.3 Sensitivity of the Monte Carlo Results

Careful analysis of any Monte Carlo simulation must include a discussion of the simulation error itself. In a world of unlimited resources, the simulation error can be driven down to negligible scales by increasing the number of replications, in our case increasing the number of simulated economies, $E$, from 1000 to several million, and increasing the fans, $J$, used in the calculation of each economy’s market price from 1000 to several million. This chapter’s Monte Carlo experiment involves a simulation of a numerical integration (to calculate market prices) as well as the method of Donaldson and Kamstra (1996) which is also a numerical integration (to calculate the DK prices). The scale of the simulation quickly reaches frightening proportions as we increase the number of replications (economies), $E$, or fans, $J$, in the numerical integrations. The choice of 1000 economies, each stretching out for 47 years, and then the choice of 1000 fans used to calculate the market prices for each year of each economy, led to roughly one month of CPU time on an SUN UltraSparc II 400 MHz machine, the practical limit for this experiment given competing demands for this resource.

To determine the simulation error, we must conduct a simulation of the simulations. Unlike some Monte Carlo experiments (such as those estimating the size of a test statistic under the null) the standard error of the simulation error for most of our estimates (returns, prices, etc.) are themselves analytically intractable, and must be simulated. In order to

---

*Less than 2% of the simulations (economies) yielded DK models that were not stable, or that produced rollouts that were not stable, and were excluded from the analysis. These economies were not otherwise remarkable, and the majority of the experimental results we present are qualitatively unchanged whether these simulations are included or not.
estimate the standard error of the simulation error in estimating market prices, we estimated a single market price 1000 times, each time independent of the other, and from this set of prices computed the mean and variance of the price estimate. If the experiment had no simulation error, each of the 1000 price estimates would be identical. With the number of fans, $J$, equal to 1000 we find that the standard deviation of the simulation error is only 0.28% of the price, which is sufficiently small as to not be a source of concern for our study.*

4 RESULTS OF THE MONTE CARLO EXPERIMENTS

We now explore results based on our simulated market economies and the fundamental prices calculated from them. The first set of results focus on summary statistics for returns, dividend–price (D/P) ratios, dividend growth rates, interest rates, and risk premia produced by our simulated economies as compared to data from real-life financial markets. The second set of results focuses on tests of bubbles in the simulated economies, as well as on some related statistics.

Recall that no experiment presented here is calibrated to actual S&P 500 prices. All experiments are calibrated only to dividend growth rates and discount rates, and then prices are calculated based on the underlying fundamental factors. Great care was taken to ensure that our experiments faithfully replicated the mean, variance, cross-correlation, and auto-regressive structure of log dividend growth rates and log discount rates. If S&P 500 prices contain bubbles, we would expect that S&P 500 returns and D/P ratios would be excessively volatile relative to our simulated economies, which do not themselves contain bubbles. Furthermore, since we have no bubbles in our simulated economies, the properties of the fundamental price estimates and the ex-post rational price under the null of no bubbles can be obtained from our simulations. This allows us to investigate the size properties of tests for bubbles based on these bubble-free price estimates.

Recall from the discussion above that we have 47 years of data from the S&P 500 and that each of our 1000 simulated economies therefore also produce 47 years of data. For each of the 1001 economies under

*We further investigated the sensitivity of selected results by increasing the number of simulated economies, $E$, or the number of fans, $J$, to as large as 10,000 (restricting to 2 years the number of periods each economy was followed out), but found no qualitative changes to our experimental results. Of course, this did reduce the simulation error by a factor of square root 10, but was not of much practical gain since the simulation error was already extremely small.
investigation (one actual economy and 1000 simulated economies) we calculate the mean, standard deviation, skewness, and kurtosis of a variety of interesting variables, including dividend growth rates, interest rates, stock returns, and D/P ratios. For example, we calculate the mean return of the 47 years of the actual S&P 500, which appears in the top-left cell of the results shown in Table 1. We also calculate the mean return for each of our 1000 simulated economies, which produces 1000 mean return estimates, the distribution of which is summarized in the remaining cells of the first row of results shown in Table 1 (the \( X \)th percentile reports the return for the economy with the \( X \)th largest return).

We begin our analysis of the results in Table 1 by considering results for dividend growth rates and discount rates. Given that we calibrated the discount rates and dividend growth rates to the true data, we would expect that discount and dividend growth rates from our simulated economies would appear quite similar to the true market data. And indeed, we see

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Statistics on the S&amp;P 500 Index and Simulated Market Index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td>Mean of simulated data</td>
</tr>
<tr>
<td><strong>Statistic</strong></td>
<td>S&amp;P 500</td>
</tr>
<tr>
<td><strong>Returns</strong></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
</tr>
<tr>
<td><strong>D/P ratios</strong></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
</tr>
<tr>
<td><strong>Dividend growth rates</strong></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
</tr>
<tr>
<td><strong>Interest rates</strong></td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
</tr>
<tr>
<td><strong>Risk premia</strong></td>
<td>Mean</td>
</tr>
</tbody>
</table>
means, standard deviations, skewness, and kurtosis for our simulated interest rates and dividend growth rates very nearly identical to the true interest rate and dividend growth rate data.* Furthermore, each and every statistic for dividend growth rates and interest rates lies near the center of the distribution of the simulated series. Given that we generated dividend growth rates and interest rates to match the true data this is not surprising and is only presented as a verification check. The true challenge, which may provide some insight into the existence of bubbles in the S&P 500 data, is to see if S&P 500 returns and prices also look like returns and prices from economies that share dividend and interest rate characteristics with the true economy, but have been simulated to have no bubbles.

From the first row of Table 1 we see that, on average, our simulated economies produce mean returns very close to returns from the actual S&P 500. Indeed, the S&P 500’s return lies near the middle of the distribution of returns from our simulations. The skew and kurtosis of S&P 500 returns also fall well within the 90% confidence interval formed by the simulated economies. Only the standard deviation of market returns is (barely) outside the 90% confidence interval formed by our simulated economies. Price levels themselves can of course not be compared directly since price is a random walk and thus the unconditional mean, standard deviation, skewness, and kurtosis do not even exist. Financial studies therefore typically focus on the D/P ratio, since prices and dividends are cointegrated. From Table 1 we see that the D/P ratio of the S&P 500 is similar to our simulated data, with all statistics measured for the S&P 500 data within a 90% confidence interval formed by the simulated economies’ statistics (with the sole exception of the mean D/P ratio, which is outside the 90% confidence interval but within the 98% confidence interval, the 5 percentile range). Again, similar to returns, the S&P 500 D/P ratios do not stand out as wildly unusual, given the behavior of returns we see in no-bubble economics.

In sum, the results for dividend growth rates, interest rates, returns, and D/P ratios shown in Table 1 suggest that the behavior of the true S&P 500 stock index over the past 50 years is not out of line with how we might expect a financial market to behave if that market priced stocks in a rational, efficient, and bubble-free manner, given interest rate and dividend fundamentals. What other studies have argued to be anomalies in price behavior, such as volatile return series, seem in our study to be entirely

*For instance, the S&P 500 dividend growth rates have a slight positive skew of 0.257, while the 5 and 95 percentiles of the simulated series of dividend growth rates are −0.45 and 0.66, respectively.
consistent with prices based on dividend and interest rate fundamentals. The key difference between our investigation and previous work is our calibration to the dividend and interest rate characteristics of the economy with no simplifying assumptions when simulating market prices, returns, and various statistics and their distributions derived from these prices and returns. Conversely, previous studies, even those that have used simulation techniques, can only make (educated) guesses at what the empirical distribution of various tests and statistics might be without the simplifying assumptions imposed. To the extent that these guesses and simplifying assumptions (e.g., independent dividend innovations) are inaccurate, or to the extent that small sample properties of the financial time series are not accounted for, misleading results can be obtained and “anomalies” can be found where none truly exist. Thus, the approach we adopt leads to a conclusion opposite to many studies in the literature, starting with Shiller (1981), which conclude that stock prices are excessively volatile relative to fundamentals.

To investigate further the summary statistics reported in Table 1, we provide a series of figures that show the bivariate distributions of the economy-by-economy statistics that went to comprise Table 1. The S&P 500 economy could conceivably be relatively usual compared to the simulated economies, statistic by statistic, but have a very unusual combination (pairing) of statistics, which would become obvious from a bivariate plot of these statistics for all the economies (e.g., the standard deviation of returns and the mean of returns for the S&P 500 might each appear similar to simulated data, but the mean-standard deviation combination seen in the S&P 500 data might be very different from the mean-standard deviation combinations produced by the simulations). The plots in Figs. 1–5 display data points for each of our simulated economies and for the S&P 500, with the actual S&P 500 data represented by a large black circle and the simulated data by smaller circled points (which can, on occasion, appear in the plots to be small black dots because of several circles partially overlying each other).*

Figure 1 presents standard deviations versus means in a group of four panels for returns (Panel A), D/P ratios (Panel B), dividend growth rates (Panel C), and interest rates (Panel D). We would expect the dividend growth rate and interest rate plots to have the actual S&P 500 data in the center of the cloud of points associated with the simulated

*For presentation, data points that completely (or very close to completely) overlay each other were condensed down to a single point.
economies since we calibrated the experiments to these two quantities, and this is indeed what we observe in Fig. 1, Panels C and D. As for returns in Panel A, we see that economies generated by calibrating to only discount rates and dividend growth rates produce mean/variance pairs for returns that do not look terribly dissimilar to the S&P 500 return data. The plot of

FIGURE 1 Mean versus standard deviation.
FIGURE 2  Return mean versus risk premium.

FIGURE 3  Interest rate mean versus risk premium.
FIGURE 4  Return mean versus interest rate mean.

FIGURE 5  Return AR(1) coefficient versus ARCH(1) coefficient.
the D/P ratios means versus standard deviations in Panel B also shows that the first two moments of the S&P 500 data are not entirely out of the bounds of possibility, although the S&P 500 data do stray to the very edge of the joint distribution of means and standard deviations for the D/P ratios from the simulated economies. In other words, prices in the S&P 500 data seem higher relative to dividends than we typically see in our simulated economies. This suggests that, while the S&P 500’s D/P ratio is not low enough to claim with any degree of confidence that current market prices contain a bubble of irrational exuberance, our evidence does suggest that actual market prices do seem rather high relative to what we observe in our simulated economies. In other words, S&P 500 stock prices may indeed be suspiciously high as some people suggest, but they are not so high as to allow us to conclude that they are not being driven by fundamentals.

Turning our attention back to Table 1, in the last row of Table 1 we present equity premia from the true data and equity premia from our simulated economies. Note that since we have only one equity premium value per economy we cannot present a standard deviation, skew, or kurtosis measure for the S&P 500 data; there is only the one equity premium estimate of 5.77%. Similarly, in each of the simulated economies we have a single measure of the equity premium and thus cannot present a standard deviation, skew, or kurtosis measure for the simulated equity premia either. However, we do see from Table 1 that the risk premia possible in a simulated economy, with dividend and interest rate processes set to mimic actual dividend and interest rate processes, can range from 0 to over 8%, even if the true equity premium is below 6%. This suggests that the equity premium we see in the actual S&P 500 data could be an imprecise estimate of the market’s true underlying risk premium and leads us to wonder whether findings of so-called anomalies, such as the Mehra and Prescott (1985) “equity premium puzzle,” are really puzzling, or whether such findings are simply small sample flukes.*

To examine further the equity premium issue, we plot in Fig. 2 the relationship between stock returns and the equity premium. Recall that the actual S&P 500 data are represented by a large black circle and the simulated data by smaller circled points. Although the high equity premium of the last half-century has been much discussed, and argued to be excessive, we see that the distribution of possible equity premia and

*For an initial identification of the equity premium puzzle see Mehra and Prescott (1985). This seminal work spawned a significant body of research on the equity premium puzzle, much of which is reviewed in Campbell et al. (1997).
mean market returns over a 47-year period is quite spread out and that the
S&P 500's experience is not terribly unusual, at least if one allows discount
rates and dividend growth rates to be autocorrelated random variables as
we do. In fact, in a bubble-free economy with fundamentals (i.e., dividend
growth rates and discount rates) calibrated to match the true economy, it
is possible to observe measured equity premia very nearly equal to mean
returns, even when we would expect that over hundreds of years the mean
return would converge to 11% (the mean return of the S&P 500 over the
past 47 years) and the mean equity premium to 6% (the equity premium
of the S&P 500 over the past 47 years). The variance of the S&P 500
discount rates and market returns is so large that there is very little we can
draw from the so-called equity premium puzzle in the short time periods
available to us in the true economy.

Some authors have argued that the equity premium is “too large,”
not because stock returns are too high, as originally argued, but rather
because risk-free interest rates are too low (Kocherlakota, 1996). Figure 3
therefore presents mean interest rates versus the equity premia. Again we
find the experience of the true data over the last 47 years to be
unremarkable relative to our simulated economies. It is not uncommon to
find in our simulations very low mean interest rates and high risk premia.
In fact, low average interest rates are virtually a prerequisite to finding
high equity premia, as the sum of the two is restricted to add up
(approximately) to the mean return, and the mean return rarely strays
above 14% over a 47-year period, according to our simulations.

For completeness, in Fig. 4 we present mean returns versus interest
rates. This combination shows the actual S&P 500 data to have a relatively
high return relative to the interest rate, but even in this case the S&P 500
data is within the envelope of the simulated economies.

For all the data and statistics we explore, including results on
returns, D/P ratios, equity premia, etc., the S&P 500 data do not seem to
be wildly unusual among economies we have generated. These economies
have been generated to be bubble-free and calibrated to the S&P 500
dividend and discount rate experience. Taken together, this evidence
therefore suggests that the S&P 500 index does not contain price bubbles.
Given this evidence that the stock market does not contain bubbles, why is
the financial econometric literature replete with tests for bubbles that in
general support the argument that stock prices are excessively volatile? It is
to this question that we now turn.

Table 2 presents the results of four common statistical tests for stock
price bubbles, presented in rows 1–4 of the table, with the columns
indicating the pricing method (the ex-post rational price and four different
fundamental price estimates) the test was based on. Each test for bubbles
was applied at the 10% significance level for each pricing method (where applicable). In Table 2 we present the actual rejection rates across our simulated economies for each of these tests and for each of the fundamental price estimation techniques. If each test is well-behaved, i.e., has proper size, and if these fundamentals estimation techniques are reliable guides to the true fundamental value, then we would expect approximately 10% rejection rates in each entry for the first four rows of the table, as all these tests are being applied under the null of no bubbles.*

From the first four rows of Table 2 we see that the most commonly applied tests for bubbles, the Camerer test and the two MRS tests, are

<table>
<thead>
<tr>
<th>Statistic or p-value</th>
<th>Ex post</th>
<th>Gordon growth</th>
<th>Gordon additive</th>
<th>Gordon geometric</th>
<th>DK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camerer test a</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>1.00000</td>
<td>0.59021</td>
</tr>
<tr>
<td>MRS1 test a</td>
<td>—</td>
<td>0.28746</td>
<td>0.06505</td>
<td>0.24873</td>
<td>0.21305</td>
</tr>
<tr>
<td>MRS2 test a</td>
<td>—</td>
<td>0.31091</td>
<td>0.32314</td>
<td>0.35474</td>
<td>0.27013</td>
</tr>
<tr>
<td>Campbell–Shiller test</td>
<td>0.31600</td>
<td>0.10296</td>
<td>0.02446</td>
<td>0.13354</td>
<td>0.21101</td>
</tr>
<tr>
<td>Mean % Error b</td>
<td>0.07682</td>
<td>0.08475</td>
<td>0.17549</td>
<td>0.00187</td>
<td>— 0.06629</td>
</tr>
<tr>
<td>% Error (σ)</td>
<td>0.17728</td>
<td>0.17842</td>
<td>0.34676</td>
<td>0.17842</td>
<td>0.15557</td>
</tr>
</tbody>
</table>

*a p-Values are from a paired-sample t-test, applied as follows. Let \(x_{1,t}\) and \(x_{2,t}\) be the two series we are interested in testing for difference in variance, let \(x_1\) and \(x_2\) be the sample means of \(x_{1,t}\) and \(x_{2,t}\), respectively, and define \(z_t = (x_{1,t} - x_1)^2 - (x_{2,t} - x_2)^2\), i.e., the difference in two paired samples. Then, under standard regularity conditions, a test of the null-hypothesis that the first moment of \(z_t\) is greater than 0 can be interpreted as a test of \(x_{1,t}\) having greater variance than \(x_{2,t}\), and implemented as a standard paired-sample t-test for the mean of \(z_t\) being greater than 0. The test statistic has an asymptotic normal distribution under fairly weak conditions, permitting non-normality as well as some dependence and heterogeneity [see Lehmann (1975) and Gastwirth and Rubin (1971)].

bPercentage error calculated as price forecast minus market price all divided by price forecast.

*The rejection rate we calculate with these simulations is a random variable, centered on 10% with a standard deviation derived from the binomial distribution, equal to the square root of 0.10 times 0.90 divided by the number of simulated economies, or 0.0095 approximately. Thus, we would expect the rejection rates to differ from 10% by no more than approximately 0.02 at the 5% level of significance.
extremely unreliable. If the tests have correct size, the $p$-value entries in Table 2 would all be 0.10. However, in all but one case the empirical $p$ values far exceed their theoretical values, which implies that these tests typically grossly over-reject. In other words, common tests for stock price bubbles find bubbles where none actually exist. Indeed, the Camerer test finds bubbles 100% of the time using ex-post and Gordon fundamental price series and 59% of the time using the DK fundamental price series, even though the market price series does not contain bubbles. Only the Campbell–Shiller test procedure (the Dickey–Fuller unit root test) comes close to having the correct size.

The last two rows of Table 2 report results on the fundamental pricing error, defined as the fundamental price minus the simulated market price, all divided by the simulated market price. Of all the fundamentals estimation techniques, the DK procedure is the most reliable forecaster of market prices, with the smallest forecast error variance and among the smallest mean errors. The superiority of the DK procedure over other simpler methods is not very great, however. All the fundamental price estimation methods we employ suffer from the same basic problem, which is the difficulty of estimating key parameters (the mean dividend growth rate and discount rate in the case of the basic Gordon growth method, the parameters of the ARMA process for discounted dividend growth rates in the case of the DK method, etc.). Clearly, based on the results of Table 2, there is scope for improvement in fundamentals estimation techniques.

Finally, we explore derived statistics from our experiments, including statistics such as the sample autocorrelation at one lag in market stock returns and the ARCH effect in stock returns. The fashion in which our data are constructed should lead to returns that are independent and identically distributed, but the distribution of the sample autocorrelation at one lag and the ARCH(1) coefficient estimate may not be well approximated with the usual asymptotic distributions. Figure 5 presents the universe of pairs of AR(1)–ARCH(1) coefficient estimates from our simulated economies’ time series of returns and also plots the coefficient estimates from the S&P 500 returns data. From Fig. 5 we see that, once again, the S&P 500 data do not appear wildly at odds with the cloud of data points from our simulated economies. Furthermore, we see from our simulated economies that by random chance we can find economies that exhibit extremely strong annual ARCH effects, with the ARCH(1) coefficients achieving values ranging from −0.4 to 0.6, and equally strong autoregressive effects, with positive and negative magnitudes of roughly 0.50 on the AR(1) coefficient, even though neither of these effects are actually present in the data.
5 CONCLUSIONS

The purpose of this chapter has been to use advanced simulation and Monte Carlo techniques to learn more about the behavior of fundamental stock prices and the properties of statistical tests commonly used to test for price bubbles in stock market data. We produce simulated stock price series based on dividend growth and discount rate fundamentals so as to simulate rational, efficient, and bubble-free economies that share dividend and interest rate characteristics with the true economy. While much conventional wisdom suggests that stock prices are “too volatile” and may be influenced by the “irrational exuberance” of traders, our evidence does not support this conclusion. On the contrary, we find that actual stock prices do not behave significantly differently from our simulated prices from rational, efficient, and bubble-free economies. We also find that traditional tests for excess volatility and bubbles over-reject the null of no bubbles in samples of the size traditionally employed in the literature. Indeed, most tests we investigate find overwhelming evidence of price bubbles even through we have constructed that data to have no bubbles. In other words, we demonstrate that traditional tests for bubbles frequently find bubbles in data that do not in fact contain bubbles.

The key difference between our investigation and previous work is our calibration to the dividend and interest rate characteristics of the economy with no simplifying assumptions when simulating market prices, returns, and various statistics and their distributions derived from these prices and returns. Conversely, previous studies, even those that have used simulation techniques, can only make (educated) guesses at what the empirical distribution of various tests and statistics might be without the simplifying assumptions imposed. To the extent that these guesses and simplifying assumptions (e.g., independent dividend innovations) are inaccurate, or to extent that small sample properties of the financial time series are not accounted for, misleading results can be obtained and “anomalies” can be found where none truly exist. While more work remains to be done in this area, we believe that the results of this chapter at least raise some concerns with some previously employed approaches and suggest an avenue for further investigation.

ACKNOWLEDGMENTS

We thank Lisa Kramer for helpful comments and the Social Sciences and Humanities Research Council of Canada for financial support. The
views expressed are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

REFERENCES


Estimating and Testing Fundamental Stock Prices


12

Neural Networks: An Econometric Tool

Johan F. Kaashoek
Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands

Herman K. van Dijk
Econometric Institute, Erasmus University Rotterdam, Rotterdam, and Tinbergen Institute, Rotterdam, The Netherlands

1 INTRODUCTION

In recent decades one witnesses a substantial increase in the interest of econometricians for nonlinear models and methods. This is due to: (1) advances in the processing power of personal computers, (2) increased and successful research on algorithms for fast numerical optimization methods, and (3) the availability of large data sets. One of the nonlinear models that received much attention from applied researchers is a neural network, also known as a neural net. The basic idea behind a neural net is the tremendous data-processing capability of the human brain. Human brains consist of an enormous number of cells, labeled neurons. These neurons are connected and signals are transmitted from one cell to another through the connections. These connections are, however, not all equally strong. When a signal is transmitted through a strong connection it arrives more strongly in the receiving neuron. One may argue that there is a particular...
weight associated with each connection, which varies with the strength of the connection. Neurons may also receive signals from outside the brains. There are then transformed within the brains and returned to the outside world. The whole structure of signal processing between many (unobserved) cells can be described by a particular mathematical model that is therefore known as an artificial neural network model. For convenience we delete, henceforth, the qualification artificial. A more detailed description of the analogy between the mathematical neural network models and the working of the human brain is given by, e.g., Simpson (1990).

Neural networks are used in many sciences like biology, informatics, and econom(etr)ics. Within the latter field neural nets are, in particular, applied for the description and prediction of complex data patterns in economic time series. The field is very extensive and empirical illustrations are many. This chapter is not intended to give a complete survey. Instead, we start with a brief introduction on neural nets and their flexibility. Our focus is on the following two applications of neural network analysis with the aim of showing that neural nets are a convenient econometric tool:

1. Recovery of the unobserved dynamics; in particular, stability of a nonlinear system from a low-dimensional data set.
2. Specification of a neural network where a time-varying component is included.

In the first topic a neural net is used to recover the dynamic properties of a nonlinear system, in particular, its stability, by making use of the Lyapunov exponent. We use one simulated series from a structurally unstable chaotic model and some data from real exchange rates to illustrate the methods. Second, a two-stage network is introduced where the usual nonlinear model is combined with time transitions that may be handled by neural nets. The connection with time-varying smooth transitions models is indicated. The procedures are illustrated on a time-varying Philips curve using U.S. data from 1960–1997. We discuss connections with the existing literature, but refer for a general introduction to neural network to Hertz et al. (1991) and Bishop (1995) and the references cited therein.

2 A SIMPLE INTRODUCTION TO NEURAL NETWORKS

There exist many classes of neural networks, see, e.g., Hertz. et al. (1991) and Bishop (1995). In this chapter we restrict attention to a simple class that is known as the three-layer feed forward neural network, also labelled the Rumelhart–Hinton–Williams multilayer network after Rumelhart et al.
(1986). For expository purpose we describe and interpret this network as a generalization of the well known linear model from basic econometrics, see, e.g., Theil (1971, Chap. 3). Suppose that the cells of the network are partitioned into particular groups or layers and suppose further that there exist three such layers: the “input” layer, the “hidden” layer, and the “output layer.” The cells of the input layer correspond to the “regressors” or “explanatory variables” in the standard linear regression model. The cell in the output layer correspond to the dependent variables in the linear model. The hidden layer contains cells that transmit the signals from the input layer to the output layer. These cells may be interpreted as unobserved components built into the linear model. A graph of a neural network with three cells in the input layer, two cells in the hidden layer and two cells in the output layer is shown in Fig. 1.

The network transmits signals as follows. A weighted sum of the signals of the input cells are sent to the hidden layer cells. Within the cells of this layer the values of the signals received are transformed by a so-called “activation function.” A weighted sum of the transformed signals is then sent to the cells of the output layer. We note that the weights in the neural network correspond to unknown parameters in the linear model.

Henceforth, we make use of the following (standard) notation. A neural network with \( I \) cells in the input layer, \( H \) cells in the hidden layer, and \( O \) cells in the output layer is denoted as \( nn(I,H,O) \). In Fig. 1 the network is given as \( nn(3,2,2) \).

**FIGURE 1** Graph of a neural network.
Next, we discuss the mathematical structure of a neural net. We make use of the following notation for cells, signals, and weights:

- \( i \) : index of input cells, \( i = 1, \ldots, I \)
- \( h \) : index of hidden layer cells, \( h = 1, \ldots, H \)
- \( j \) : index of output cells, \( j = 1, \ldots, O \)
- \( g(\cdot) \) : activation function
- \( x_i \) : value of input cell \( i \)
- \( y_j \) : value of output cell \( j \)
- \( a_{ih} \) : weight of the signal from input cell \( i \) to hidden cell \( h \)
- \( b_h \) : constant input weight for hidden cell \( h \)
- \( c_{jh} \) : weight of the signal from hidden cell \( h \) to output cell \( j \)
- \( d_j \) : constant weight for output cell \( j \)

The value of the signal that arrives in hidden cell \( h \) is given as

\[
\text{Input for hidden cell } h = \sum_{i=1}^{I} (a_{ih}x_i) + b_h
\]  

(1)

The hidden cell \( h \) transforms the value of this signal with the activation function \( g(\cdot) \) as follows:

\[
\text{Output from hidden cell } h = g\left(\sum_{i=1}^{I} (a_{ih}x_i) + b_h\right)
\]  

(2)

where the activation function is a monotone increasing and bounded function given as

\[
g(x) = \frac{1}{1 + e^{-x}}
\]  

(3)

which is the well known logistic function, defined on the interval \([0,1]\). A particular value of the logistic function indicates the extent to which a hidden cell is activated. The logistic function has attractive properties such as that the derivative is equal to \( g(x)(1 - g(x)) \). The graph of the function is given in Fig. 2.

Other choices of the activation function are other monotone squashing functions as the \( \arctan \) and \( \tanh \) functions and the \( \cos \) squashing function, see e.g., Hertz. et al. (1991).

Next, the value \( y_j \) of the output cell \( j \) is given as the weighted sum of the output of the hidden cells. It is equal to

\[
y_j = \sum_{h=1}^{H} c_{jh}g\left(\sum_{i=1}^{I} a_{ih}x_i + b_h\right) + d_j
\]  

(4)
In matrix notation one can write:

\[ y = CH + d \]  \hspace{1cm} (5) 

\[ \mathcal{H} = G(xA + b) \]  \hspace{1cm} (6)

where

\begin{align*}
    x & \in \mathbb{R}^I \\
    A & = (a_{ij}) \text{, } I \times H \text{ matrix} \\
    \mathcal{H} & = (h_1, \ldots, h_H) \text{, the vector of hidden cell outputs} \\
    G : \mathbb{R}^H & \to \mathbb{R}^H \text{ is the vector function, given by} \\
    G(v) & = [g(v_1), \ldots, g(v_H)]' \\
    C & = (c_{jk}) \text{, } O \times H \text{ matrix} \\
    y & \in \mathbb{R}^O \\
    b & \in \mathbb{R}^H \\

    \end{align*}

The neural network \( mn(\cdot) \) describes the situation at one moment in time and it indicates a deterministic relation. By adding a subscript \( t \) and an error term \( \varepsilon \) one obtains the system:

\[ y_t = CH_t + d + \varepsilon_t \]  \hspace{1cm} (7) 

\[ \mathcal{H}_t = G(x_tA + b) \]

Figure 3 is a representation of an \( mn(3, 2, 1) \). The mathematical specification of this model is equal to

\[ y_t = d + \frac{c_1}{1 + e^{-a_{11}y_{t-1} - a_{12}y_{t-2} - h_1}} + \frac{c_2}{1 + e^{-a_{21}y_{t-1} - a_{22}y_{t-2} - h_2}} \]  \hspace{1cm} (8)
We note that this model is closely related to a threshold autoregressive model: for details see Granger and Teräsvirta (1993) and Van Dijk (1999).

2.1 Flexibility of Neural Networks

The flexibility of three-layer feed forward neural nets is well documented. It is summarized by its so-called "universal approximation" property. Most of this approximation theory starts with Kolmogorov’s representation theorem (1957). This provides the background for the Hecht-Nielsen article (1987). From the point of view of a neural network user, the Kolmogorov theorem provides, however, a justification for the existence of approximations in the reverse way. That is to say, the number of layers and cells are given but not the functional form of the one-dimensional activation functions. In a neural network one encounters the opposite case: the activation functions \( g \) are given (to some extent) but, at least, the number of hidden layer cells is unknown.

The articles of Gallant and White (1989) (with the revealing title: "There exists a neural network that does not make avoidable mistakes"), Cybenko (1989), Funahashi (1989), Hecht-Nielsen (1989, 1990) and
Hornik et al. (1989). Lesho et al. (1993) and White (1989), provide the theoretical background for the statement: “A (three layer) feed forward network is an universal approximator.” This general statement should be interpreted in the sense that any square integrable function can be approximated arbitrary close in $L_2$ norm.

Further, the articles of Hornik et al. (1990) and Gallant and White (1992) extended the approximation capabilities of the network to the derivative of a function.

2.2 Estimation of Parameters of Neural Networks

A generally accepted optimization principle is to minimize the norm of

$$\min_{\theta} \sum_{t} \|y_t - \hat{y}(\theta)\|^2$$

where $\|\|$ is the Euclidean norm. For the case of a neural net it follows that one minimizes the criterion function:

$$\min_{A, b, C, d} \sum_{t} \|y_t - CG(x_tA + b) - d\|^2$$

Well known methods for numerical optimization are the simplex method and the BFGS method, which is a gradient method (see Press et al., 1988).

Initial values for the parameters $A$, $b$, $C$, and $d$ are chosen randomly from the uniform distribution on $[-0.5, 0.5]$. We emphasize that the logistic function is insensitive for large and small values. Therefore one should scale down the range of data before processing through the network. Since the output range of a hidden cell lies between 0 and 1, an appropriate range for the data should be $[0.1, 0.9]$.

2.3 Determining the Size of a Neural Network

Neural nets are flexible, but the price of increased flexibility is the danger of “overfitting.” This statement may be explained as follows.

In empirical econometric models one assumes that an observed economic time series consists of a part that can be explained and a part that is labelled unexplained or “residual noise.” With “overfitting” this noise is also “fitted.” One then obtains a wrong picture of the real data-generating process and the quality of the forecasts may be badly affected. “Overfitting” with neural nets may occur by increasing the number of hidden cells, which increases the number of parameters, without increasing the number of explanatory variables or inputs. Because of this possibility neural nets are more sensitive to “overfitting” than other classes of models like autoregressive models. Therefore, it is important to develop methods
that determine the optimal size of a neural net. Pruning methods apply to the reduction of large neural networks to smaller ones. Two methods can be distinguished: weight (interconnection) reduction or node reduction. Examples can be found in Hertz et al. (1991) and Bishop (1995) (see also Mozer and Smolensky, 1980).

Below we summarize a descriptive method to reduce the size of a neural net. Descriptive methods are useful for exploratory data analysis. This type of analysis is more and more needed since in recent decades large data sets for economic variables have become available. In these data, complex patterns of the economic variables may occur. For more details we refer to Kaashoek and van Dijk (1998).

2.3.1 Pruning a Network: Incremental Contribution Method

The method we follow is labeled incremental contribution method.” It looks for each cell separately how much the specific cell contributes to the overall performance of the network. When this contribution is considered to be low then such a cell is a candidate for excluding from the existing network (and all its connections). Re-estimating the reduced network may confirm this exclusion. To measure the contribution of a cell we look at two quantities.

First, the square of the correlation coefficient ($R^2$) between $y$ and $\hat{y}$, the neural network output where

$$R^2 = \frac{(\hat{y}^2 y)^2}{(\hat{y} y)(\hat{y} y)}$$

(11)

and $y$ as well as $\hat{y}$ are taken in deviation of the means. The procedure applies to hidden layer cells as well as to input layer cells, but here we restrict ourselves to hidden layer cells.

The contribution of cell $h$ can now be measured by leaving out this cell and its connection from the network, and again calculating the square of the correlation coefficient; we denote network estimates with cell $h$ left out by $\hat{y}_{-h}$, and the corresponding $R^2$ is defined as

$$R^2_{-h} = \frac{(\hat{y}_{-h} y)^2}{(\hat{y} y)(\hat{y}_{-h} \hat{y}_{-h})}$$

(12)

The incremental contribution $R^2_{incr} (i)$ is now given as

$$R^2_{incr}(h) = R^2 - R^2_{-h}$$

(13)

In the group of hidden layer cells, cells with a low $R^2_{incr}$ are candidates for exclusion.
The second quantity involves the idea of principal components (Malinvaud, 1970; Theil, 1971). Let again $\hat{y}_h$ be the network output with exclusion hidden layer cell $h$. Construct the vector $e_{-h}$ of residuals:

$$e_{-h} = y - \hat{y}_{-h}$$

and the matrix $E_{-H}$:

$$E_{-H} = (e_{-1, \ldots, e_{H-H}})$$

The matrix $E_{-H}E_{-H}/T$ is an estimate of the covariance matrix of the $e_{-h}$'s. The principal component of $E_{-H}E_{-H}$ is the eigenvector at maximal eigenvalue in the absolute sense. Hence, the principal components provide a linear combination of $(e_{-1, \ldots, e_{-H}})$, which explains the largest part of the variance.

Which fraction is explained by each of the eigenvectors is given by the relative weight $w_i$ with

$$w_h = \frac{\lambda_h}{\sum_{h=1}^{H} \lambda_h}$$

where $\lambda_h$ are the eigenvalues of $E_{-H}E_{-H}$. In the case of the principal component, $\lambda_H$ is the largest eigenvalue.

Now one can look at the components of the principal component itself: cell with low incremental contribution will have a low (in the absolute sense) coefficient in the eigenvector composing the principal component. The exclusion of those cells will cause a relatively small increase in the residuals; these cells are again candidates for exclusion.

Finally, one may apply a graphical analysis. Assuming the graph of $\hat{y}$ fits well with the graph of $y$, the graph of $\hat{y}_{-h}$ may differ less from the graph of $y$ for those cells with a "low contribution."

The procedure involves the contributions of one cell only. It is a "feature" of neural networks that sometimes pairs of cells do have a similar contribution with the output of the cells having reverse sign. Such a "behavior" can be detected by graphical analysis and by observing that in the principal component analysis, explained above, such types of cells do have (almost) equal coefficients. In that case one has to look at the incremental contribution of both cells together.

We end this section with a remark on the descriptive nature of our procedure. As stated before, large data sets have become available in the economic sciences. Important examples are data set on household behavior and labour employment at the individual level. Further, data sets on financial variables in the stock and exchange rate markets. This appears as only a beginning. The new scanner data in marketing constitute great and fascinating challenges in data analysis. Because of its flexibility, we believe
that a “pruned” neural network may be a useful instrument for exploratory data analysis in order to find possibly complex patterns of economic variables.

Our method is descriptive and as such has the same limitations as all techniques of data summarizations like the construction of histograms and/or the plotting of time series together with a list of summary statistics. However, using neural networks one can perform a dynamic analysis and compute the long-run properties (see Section 3). For a statistical approach to neural network analysis in econometrics, we refer to White (2000).

3 STABILITY ANALYSIS OF COMPLEX NONLINEAR SYSTEMS

A linear autoregressive system of equations of the $n$-vector of variables $x(t)$ can be written as

$$x_{t+1} = Ax_t, x_t \in \mathbb{R}^N$$  \hspace{1cm} (17)

The stability of fixed or equilibrium points depends on the eigenvalues $\lambda$ (real or complex) of the matrix $A$. Taking the absolute value or modulus of the eigenvalues, $||\lambda||$, the fixed point will be unstable if for some eigenvalue $\lambda$, $||\lambda|| > 1$, which is equivalent to $\ln||\lambda|| > 0$. The same holds for the stability of orbits or time series $x_0, x_1, \ldots$. Note that in a point $x_t \in \mathbb{R}^N$, the logarithm of the local expansion rate in the direction of a vector $v \in \mathbb{R}^N$ is given as

$$\ln \left| \frac{A^v}{||v||} \right|$$  \hspace{1cm} (18)

Lyapunov exponents are a generalization of the above concept for nonlinear systems. They are defined as the (spatial or time) means of the logarithm of local expansion rates. In the case of time means, the expansion rates are calculated in the time series $x_0, x_1, \ldots$ where in a point $x_t \in \mathbb{R}^N$, the logarithm of the local expansion rate in the direction of a vector $v \in \mathbb{R}^N$ is now given as

$$\ln \left| \frac{D_x F(x_t)^v}{||v||} \right|$$  \hspace{1cm} (19)

where $F: \mathbb{R}^N \to \mathbb{R}^N$ is the data-generating function:

$$x_{t+1} = F(x_t)$$  \hspace{1cm} (20)
and $D_s F$ is the Jacobian. Then, the Lyapunov exponent $\lambda_{v_0}$, with start direction vector $v_0$ and start value $x_0$, is defined as

$$\lambda_{v_0} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \ln \left\| D_s F(x_{t-1}) \frac{v_{t-1}}{\|v_{t-1}\|} \right\|$$

(21)

$$v_t = D_t F(x_{t-1}) v_{t-1}, \quad \|v_0\| = 1$$

(22)

The dependence of $\lambda$ on $v_0$ and on $x_0$ seems to indicate an infinite number of Lyapunov exponents. However, this is not the case. In general, there are as many Lyapunov exponents as the dimension $N$ of the system; see (Guckenheimer and Holmes, 1983). Above all, for an ergodic system, with the space mean being equal to time mean, it follows that for almost all start directions $v_0$ and for almost all initial values $x_0$, the value of $\lambda_{v_0}$ will be the largest Lyapunov exponent (Arnold and Avez, 1988; Guckenheimer and Holmes, 1983).

Since the mean of logarithms is the logarithm of the geometric mean, one can also write

$$\lambda_{v_0} = \lim_{T \to \infty} \frac{1}{T} \ln \left( D_s F(x_{t-1}) \frac{v_{t-1}}{\|v_{t-1}\|} \right)$$

(23)

$$= \lim_{T \to \infty} \frac{1}{T} \ln \| D_s F(x_0) v_0 \|$$

(24)

If some (or all) of the Lyapunov exponents are positive then one has so-called “sensitivity on initial values,” a characteristic of chaotic series. Hence, in particular, the largest Lyapunov exponent is of interest: if positive, then the series is unstable, if negative then the series is stable. This result can be applied directly if $F$, the data-generating function, is known. However, in practice, one observes only a one-dimensional, finite time series $\{x_t\}$, $t = 0, \ldots, T - 1$. So the question is how to extract from the series $\{x_t\}$ the dynamic properties, especially the value of the largest Lyapunov exponent of the original (unknown) model, Eq. (20).

Although only the series $x_t$ is given, one can use the embedding theorem of Takens (1981) to reconstruct from the one-dimensional series $x_t$ the original deterministic and smooth model. This theorem says that if the data $x_t$ have a deterministic explanation, which means that the data-generating process is smooth and deterministic, there exists a finite embedding $m$, such that the dynamic system $\Psi : \mathbb{R}^m \to \mathbb{R}^m$ given by

$$\Psi : (x_t, x_{t-1}, \ldots, x_{t-m+1}) \to (x_{t+1}, x_t, \ldots, x_{t-m+2})$$

(25)
has the same dynamical properties as the original one. Moreover, if the original system is \( N \)-dimensional then an embedding dimension \( m = 2N + 1 \) will be sufficient to have “\( \Psi \) reconstruct the original system.”

Note that the only unknown component of \( \Psi \) is the first component function \( \Psi_1 : \mathbb{R}^m \to \mathbb{R} \) with

\[
\Psi_1 : (x_t, x_{t-1}, \ldots, x_{t-m+1}) \to x_{t+1}
\]

For the other components, \( i > 1 \), yields \( \Psi_i(x_t, x_{t-1}, \ldots, x_{t-m+1}) = x_{t-i+2} \). This function \( \Psi_1 \) should be approximated, and a neural network seems to be a proper candidate to do the job because of its “universal approximation” property (Kaashoek and van Dijk, 1994).

Once \( \Psi \) is found, the Lyapunov exponents can be calculated using Eq. (23). This means the calculation of \( D_x \Psi \) in each point of the time series \( (x_t, x_{t-1}, \ldots, x_{t-m+1}) \). Note that \( D_x \Psi \) has the form of a companion matrix. However, to avoid overflow in calculating the product of \( D_x \Psi \)’s the more stable method of Eckmann–Ruelle is used. In this case, a \( QR \) decomposition of \( D_x \Psi \) is calculated at each point and the Lyapunov exponents are now simply the products of diagonal elements of the matrices \( R \) (Eckmann and Ruelle, 1985).

We apply this procedure to two data sets, one simulated and one economic time series.

### 3.1 Simulated Data Experiment

In this experiment we use simulated data. The data are generated by the model:

\[
\begin{align*}
x_{t+1} &= \alpha x_t + \varepsilon_t - 0.5 \\
\varepsilon_{t+1} &= \gamma \varepsilon_t (1 - \varepsilon_t)
\end{align*}
\]

where only the series \( x_t \) is observed. The data, called \( CH95 \), are generated with \( \alpha = 0.95 \) and \( \gamma = 4 \). Although completely deterministic, this model has some nice features:

- The series \( \varepsilon_t \) is chaotic: for initial values in \([0, 1]\), the series \( \varepsilon_t \) is bounded between 0 and 1 but has the “sensitivity of initial values” property characteristic for chaotic series.
- The model is structurally unstable (\( \gamma = 4 \)) e.g., a small change in the coefficient value 4 will cause a dynamically different data series. For instance, a \( \gamma \) value greater than 4 will for almost all initial values (between 0 and 1) generate diverging (exploding) data \( \varepsilon_t \); for values less than 4, periodic data are possible: the system is structurally unstable.
Suppose that only the one-dimensional data set $x_t$ is observed. Our goal should be to extract from this series, the dynamical properties of the original data-generating process: only Lyapunov exponents are considered.

Note that model equation (27) is equivalent to a second-order nonlinear difference equation in $x_t$:

$$x_{t+2} = ax_{t+1} + \gamma(x_{t+1} - ax_t + 0.5)(1 - x_{t+1} + ax_t - 0.5)$$

where the right-hand side is a (nonlinear) continuous function in $(x_{t+1}, x_t)$. So this case is well suited for a neural network approximation. Moreover, since the neural network is intrinsically nonlinear, the chaotic aspect of the data could be grasped by such an approximation while a simple linear autoregressive model would certainly fail as an “accurate” approximation.

The original data generating model, Eq. (27), has two Lyapunov exponents, which can be calculated analytically. In a point $(x_t, x_{t-1})$, the Jacobian of the system function $F$ is given by

$$\begin{pmatrix} 0.95 & 1 \\ 0 & 4(1 - \varepsilon_t) \end{pmatrix}$$

(29)

Since the Jacobian $D_x F(x_0, \varepsilon_0) = \prod_{i=1}^T D_x F(x_{i-1}, \varepsilon_{i-1})$, one has

$$D_x F(x_0, \varepsilon_0) = \begin{pmatrix} 0.95^T & a_{12} \\ 0 & a_{22}(T) \end{pmatrix}$$

(30)

where $a_{22}(T) = \prod_{i=1}^T 4(1 - \varepsilon_{i-1})$. It is obvious that $D_x F(x_0, \varepsilon_0)$ has two eigenvalues: namely $0.95^T$ and $a_{22}(T)$. With $v_0 = (1, 0)$, $D_x F(x_0, \varepsilon_0)v_0 = [(0.95)^T, 0]$, which results in a Lyapunov exponent value of $\ln(0.95)$. It is well known that the system $\varepsilon_t = 4\varepsilon_{t-1}(1 - \varepsilon_{t-1})$ has Lyapunov exponent $\ln(2)$. Since $\ln(2) > \ln(0.95)$, for any start vector $v_0 \neq (1, 0)$, the resulting Lyapunov exponent will be $\ln(2)$. Writing both Lyapunov exponents in base 2 logarithm, the values will be $\ln(0.95)/\ln(2) \approx 0.074$ and 1.

Before starting with the neural network computations of the function $\Psi_1$ see Eq. (26), first the scatter diagram $(x_{t-1}, x_t)$ of the series $CH95$ is given. The sample size $T$ is 200. Since $\varepsilon_t$ is bounded between 0 and 1, the graph of $(x_{t-1}, x_t)$ lies between the lines $y = 0.95x + 0.5$ and $y = 0.95x - 0.5$. (See Fig. 4.)

As said above, in order to extract from the one-dimensional observed data $\{x_t\}$, the dynamic properties, especially the value of the largest Lyapunov exponent of the original model, Eq. (27), only the function:

$$\Psi_1(x_t, x_{t-1}, \ldots, x_{t-m+1}) = x_{t+1}$$

should be approximated by a neural network. This implies a choice for the value of $m$, or in neural terms, the size of the input layer. Since the original system has dimension 2, based on Takens embedding theorem, it will be
(more than) sufficient to take as neural network input variables \(x_t, x_{t-1}, \ldots, x_{t-5}\): adding an additional constant, the dimension of the input layer will be 6.

The initial number of hidden layers will be 5, while the output layer has only one cell, the target value being \(x_{t+1}\). Before processing the data through the neural network, the data are scaled down to a range of \([0.1, 0.9]\). Linear scaling does not affect the value of the Lyapunov exponents. Initial parameter values are randomly chosen from \([-0.5, 0.5]\).

To summarize the performance of this network, the quantities:

\[
R^2 = \frac{(y - \hat{y})^2}{(y) \hat{(y) (y)}} \quad (31)
\]

\[
MSSR = \frac{1}{T} (y - \hat{y}) (y - \hat{y}) \quad (32)
\]

\[
SIC = \ln(MSSR) + \frac{n_p}{2T} \ln(T) \quad (33)
\]

are calculated. In Table 1 the results for the \(mn(6,5,1)\) are reported.*

*All values reported are rounded-off at the last reported digit.

**FIGURE 4** Scatter diagram of series CH95.
In Table 2 the incremental contributions of the hidden layer nodes are given: both $R^2$ and principal component vector (row with label PrincComp. in Table 2 indicate that two hidden layer nodes can be removed.

The incremental contributions of inputs are shown in Table 3: the inputs $x_{t-4}$, $x_{t-3}$, $x_{t-2}$ should be removed. Applying node removal, the network is reduced to two inputs plus a constant and three hidden layer nodes (plus constant). This $nn(3, 3, 1)$ performs as well as the larger network (see Table 4). Both input nodes do have the same contributions: no further reduction at this level is applied.

The incremental contributions of hidden layer nodes is shown in Table 5. Based on the principal component vector (with weight 99%) hidden node 1 could be removed (see again Table 4 for the results on this network). The performance is just slightly worse compared to the larger

### Table 1

<table>
<thead>
<tr>
<th>Network</th>
<th>$R^2$</th>
<th>MSSR</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$nn(6, 5, 1)$</td>
<td>1.00</td>
<td>$1.4 \times 10^{-9}$</td>
<td>-9.70</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Hidden cell</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_{incr}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.517</td>
<td>0.880</td>
<td>0.940</td>
</tr>
<tr>
<td>PrincComp.</td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.081</td>
<td>-0.733</td>
<td>0.675</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Input cell</th>
<th>$x_{t-4}$</th>
<th>$x_{t-3}$</th>
<th>$x_{t-2}$</th>
<th>$x_{t-1}$</th>
<th>$x_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_{incr}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.124</td>
<td>0.241</td>
</tr>
<tr>
<td>PrincComp.</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.613</td>
<td>0.790</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Network</th>
<th>$R^2$</th>
<th>MSSR</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$nn(3, 3, 1)$</td>
<td>1.00</td>
<td>$1.5 \times 10^{-9}$</td>
<td>-10.04</td>
</tr>
<tr>
<td>$nn(3, 2, 1)$</td>
<td>1.00</td>
<td>$6.3 \times 10^{-9}$</td>
<td>-9.31</td>
</tr>
</tbody>
</table>
network. No further reduction is applied. So we end up with a network of two inputs $x_{t-1}, x_t$ (plus constant) and two hidden layer nodes.

Based on this $nn(3, 2, 1)$ network, the Lyapunov exponents of $\Psi$ are calculated. This can be done in two ways: either along the actual data $(x_t, x_{t-1})$, or along a series $(\hat{x}_{t-1}, \hat{x}_t)$, where $\hat{x}_t$ is a series generated by the neural network function $nn(3, 2, 1)$:

$$\hat{x}_{t+1} = nn(3, 2, 1)(\hat{x}_{t-1}, \hat{x}_t)$$  \hspace{1cm} (34)

The series $\hat{x}_t$ is the dynamic forecast given some initial values $\hat{x}_0 = x_0$, $\hat{x}_1 = x_1$; here and in the following, such a series will be denoted as orbit in contrast to the actual data series $\{x_t\}$.

If the function $\Psi$ is indeed a proper approximation of the original model then one should expect that the dynamic properties along the actual data and along the orbit data should be similar.

The results for the Lyapunov exponents are given in Table 6; they are in both cases near the theoretical values. Note that for the orbit data the largest Lyapunov exponent is greater than 1; this would correspond in the original model, Eq. (27), with a coefficient of $\varepsilon_t (1 - \varepsilon_t)$ larger than 4: for almost all initial values, the series $\varepsilon_t$ would diverge. However, the orbit data $\hat{x}_t$ converge to a large amplitude periodic pattern: in Fig. 5 the graph of $(t, \hat{x}_t)$ for the $t = 1, \ldots, 150$ (left panel in Fig. 5) and for $t = 1, \ldots, 300$ (right panel in Fig. 5) is compared to the graph of the actual data $(t, x_t)$. Since the original system is chaotic, one should not expect that both series, actual and orbit data, fall together but one should expect that the graph of both series shows a similar pattern; even in this case (and also in the larger

\begin{table}
\centering
\caption{Contribution of Hidden Layer Nodes in $nn(3, 3, 1)$ Network}
\begin{tabular}{|l|c|c|c|}
\hline
Hidden cell & 1 & 2 & 3 \\
\hline
$R^2_{inc}$ & 0.696 & 0.890 & 0.943 \\
PrincComp. & 0.114 & -0.745 & 0.657 \\
\hline
\end{tabular}
\end{table}

\begin{table}
\centering
\caption{Lyapunov Exponents $\gamma = 4$}
\begin{tabular}{|l|c|c|}
\hline
Series & Actual $x_t$ & Orbit $\hat{x}_t$ \\
\hline
Lyapunov exponent & 0.9951 & 1.0155 \\
Lyapunov exponent & -0.0735 & -0.0729 \\
\hline
\end{tabular}
\end{table}

*The initial values are taken from the actual data.
FIGURE 5 Actual (scaled) data C/365, $y = 4$ and orbit data (continuous line) generated by neural network $nn(3, 2, 1)$. Left panel: time index from 1 to 150; right panel: time index from 1 to 300.
networks obtained before) the orbit trajectory finally deviates essentially from the original pattern; see Fig. 5. Although initially the orbit data stay in the range [0.1, 0.9] of the original data, (see left panel of Fig. 5), the orbit data converge to a large amplitude (~50) periodic pattern (see the right panel of the same figure).

The reason for the deviations of orbit data from the original pattern is to be found in the structural instability of the system [Eq. (27)] with $\gamma = 4$. If $\gamma = 3.95$ then still a chaotic series will be generated; however, the system itself is structurally stable. Approximating those data by a neural network, a network similar in size as before is found. In Fig. 6 again actual and orbit (even extended in time beyond the range of the original data) are shown. Now the orbit data, the dynamical forecast of given value $x_0$, “behaves” like the original data. The same holds for the Lyapunov exponents (see Table 7). Note that, for $\gamma = 3.95$, no analytic value of the

*The initial values are taken from the actual data.
The largest Lyapunov exponent is available. Analytical values of Lyapunov exponents can be calculated if the data-generating process is ergodic and if the space distribution of the data is known; see, for more details, Arnold and Avez (1988). For the process \( t + 1 = 3.95 \frac{t}{C0} \) no analytical expression is known for the distribution of the data \( t \). The second Lyapunov exponent will still be \( /C24/C0 0.074 \).

Finally, we look how the network \( nn(3, 2, 1) \) has approximated the original data. For both cases, \( /C13 4, /C13 3.95, \) none of the parameters \( A, b, C, \) and \( d \) are very large. For instance in the case of \( /C13 4, \) the input vector of hidden nodes \( xA + b, \) is given as

\[
(xA + b)_1 = -3.1557x_{t-1} + 3.3197x_t - 1.3688 \\
= 3.3197(-0.9506x_{t-1} + x_t - 0.4123) \\
(xA + b)_2 = -0.7453x_{t-1} + 0.8095x_t + 0.0516 \\
= 0.8095(-0.9207x_{t-1} + x_t + 0.0637)
\]

Note that reported parameter values are estimation results using scaled data.

### 3.2 Nonlinear Trends in Real Exchange Rates

The economic time series are monthly observations of the natural logarithm of real exchange rates. We report those between the yen and the dollar (January 1957 to March 1998). This series is denoted by \( JPUS \) (see Fig. 7). Since the original data process is unknown, a proper embedding dimension \( m \), e.g., delay vector \( (x_t, \ldots, x_{t-m+1}) \), is also unknown. One way out is to extract from the correlation dimension the embedding dimension (Grassberger and Procaccia, 1983; Kaashoek and van Dijk, 1991).

Another approach would be to start with a rather large network and prune this network until further reduction would essentially corrupt the performance. In this case, the initial network was taken to be \( nn(6, 10, 1) \) with input variables \( (1, x_t, x_{t-1}, \ldots, x_{t-4}) \). The performance of this network, and the successively reduced networks, are summarized in Table 8. At the same time, in the column “\( \lambda \)” the largest Lyapunov exponent along the actual data set is reported while in column “\( \lambda \)” the

<table>
<thead>
<tr>
<th>Series</th>
<th>Actual ( x_t )</th>
<th>Orbit ( x_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lyapunov exponent</td>
<td>0.8618</td>
<td>0.8614</td>
</tr>
<tr>
<td>Lyapunov exponent</td>
<td>-0.0782</td>
<td>-0.0727</td>
</tr>
</tbody>
</table>
largest Lyapunov exponent along the orbit is reported. The applied reduction can be found in the column “Pruning.”

The input variables of the resulting network $nn(2, 3, 1)$ are $\{(1, x_i)\}$. In Fig. 8 the orbit based on the $nn(2, 3, 1)$ network is shown (the data are scaled down to the range [0.1, 0.9]): the orbit follows nicely the pattern of the actual data converging to a scaled value of 0.26 compatible with the unscaled value of $-0.74$.  

TABLE 8  Results of Neural Network Approximation of JPUS Data

<table>
<thead>
<tr>
<th>Network</th>
<th>$R^2$</th>
<th>MSSR</th>
<th>SIC</th>
<th>$\lambda$</th>
<th>$\hat{\lambda}$</th>
<th>Pruning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$nn(6, 10, 1)$</td>
<td>0.997</td>
<td>$0.16 \times 10^{-3}$</td>
<td>-3.880</td>
<td>0.058</td>
<td>$-0.013$</td>
<td>3 redundant hidden cells</td>
</tr>
<tr>
<td>$nn(6, 7, 1)$</td>
<td>0.997</td>
<td>$0.16 \times 10^{-3}$</td>
<td>-4.041</td>
<td>0.033</td>
<td>$-0.038$</td>
<td>2 redundant hidden cells</td>
</tr>
<tr>
<td>$nn(6, 5, 1)$</td>
<td>0.997</td>
<td>$0.16 \times 10^{-3}$</td>
<td>-4.130</td>
<td>0.048</td>
<td>$-0.143$</td>
<td>2 redundant input cells</td>
</tr>
<tr>
<td>$nn(4, 5, 1)$</td>
<td>0.997</td>
<td>$0.16 \times 10^{-3}$</td>
<td>-4.194</td>
<td>0.048</td>
<td>$-0.032$</td>
<td>2 redundant hidden cells</td>
</tr>
<tr>
<td>$nn(4, 3, 1)$</td>
<td>0.996</td>
<td>$0.18 \times 10^{-3}$</td>
<td>-4.188</td>
<td>0.076</td>
<td>$-0.034$</td>
<td>3 redundant input cells</td>
</tr>
<tr>
<td>$nn(2, 3, 1)$</td>
<td>0.996</td>
<td>$0.19 \times 10^{-3}$</td>
<td>-4.215</td>
<td>0.011</td>
<td>$-0.019$</td>
<td>3 redundant input cells</td>
</tr>
</tbody>
</table>

In all cases the orbit is stable (negative largest Lyapunov exponent) while along the actual data, the Lyapunov exponent is positive indicating an unstable series except for the final \( mn(2, 3, 1) \) network. The differences between the networks with respect to statistics are marginal. Note also that the Lyapunov exponent based on actual data still encompasses stochastic elements if present in the data itself. The smallest network is preferable.

Our empirical analysis indicates that nonlinear trends may become an important part of a modeling strategy. So far, the empirical literature on real exchange rate analysis makes mostly use of autoregressive models with possible unit root behavior; see, e.g., Schotman and van Dijk (1991) and the references cited therein, in particular, Dornbusch (1976), Frankel (1979), and Meese and Rogoff (1988).

An issue in that literature is whether or not the hypothesis of purchasing power parity (\( PPP \)) holds. Nonrejection of a unit root has led researchers to conclude that \( PPP \) does not hold. The results presented in this chapter indicate that the \( PPP \) hypothesis should be investigated on a nonlinear model.

Our empirical analysis may also be used to verify whether or not the long-run equilibrium values of several exchange rates of countries of the
European Monetary Union correspond to the values set by the European Central Bank in January 1999. For a more detailed analysis on nonlinear trends in real exchange rates of several industrialized countries using neural networks we refer to Kaashoek and van Dijk (1999).

We note that in the present chapter our interest is in an analysis of the level of the exchange rates. When one is interested in the return of exchange rates, one may perform a similar descriptive analysis on first differences. This is a topic for further research.

4 PHILLIPS CURVE

The data are monthly U.S. unemployment rates (all workers, 16 years and older), denoted by \( LHUR \), and monthly 12-period inflation rates defined by \( 100 \ln(p_t/p_{t-12}) \) at the level of the consumer price index \( p_t \); those data are denoted by \( INFR_{12} \). The unemployment rates and price indices data start at 1960 and end at November 1997.

The Phillips curve relates unemployment rates with inflation rates: a common approach (see, e.g., Sargent, 1999) is to link unemployment of one year before with current inflation. In Fig. 9 the time series of \( LHUR(-1)^* \) and of \( INFR_{12} \) are shown.

*Here, and in the following, 1-year delayed data are denoted by an \(-1\) argument.

**FIGURE 9** Time series of inflation rates \( INFR_{12} \) and unemployment rates \( LHUR(-1) \).
The Phillips curve data \(( LHUR(-1), INFR12)\) are shown in Fig. 10. Two attempts are made to model the relation between \( LHUR(-1) \) and \( INFR12 \).

First: Let \( nn(2, 6, 1) \) be the neural network with inputs a constant term and the variable \( LHUR(-1) \), consider six hidden layer cells and let the target output value be \( INFR12 \). The performance of this network, which is to fit the Phillips curve to the data of Fig. 10, is poor: \( MUSSR = 0.13, R^2 = 0.11 \). This is rather obvious because with input variable \( LHUR(-1) \), the points of Fig. 10 can hardly be considered as generated by a single valued relation (function). The graph of the neural network function \( nn(2, 6, 1)(x) \) confirms this: it seems to consists of four descending functions (all four being standard Phillips curves) with smooth transitions in between (see Fig. 11). Networks with more hidden layers show similar patterns. Although a neural network is capable of generating step-functions, it cannot of course model "multi-level relations." The graph in Fig 10 is rather to be considered as generated by some explicit time-dependent relation \((t, LHUR(-1)) \rightarrow INFR12\).

Second attempt: Referring again to Fig. 11, a time-varying approach is tempting, as in time-varying smooth transition models (see, e.g., Van Dijk, 1999). In this section, we will add an additional argument to the notation of a neural network, either \((x)\) or \((t)\) depending on whether the input variables include explicitly data \( x \) or time indices \( t \).
Suppose four time periods, and for each period a neural network approximation (of unknown size) written as $nn_i(x)$, $i = 1, \ldots, 4$ with

$$nn_i(x) = c_i g(a_i x + b_i) + d_i, \forall \ i = 1, \ldots, 4$$

then a formulation of time-varying smooth transition model could be

$$nn_1(x)g_1(t) + nn_2(x)g_2(t) + nn_3(x)g_3(t) + nn_4 g_4(t)$$

subject to some normalization, say

$$g_1(t) + g_2(t) + g_3(t) + g_4(t) = \gamma, \forall t$$

Hence, Eq. (37) can be written as

$$nn_1(x)(\gamma - g_2(t) - g_3(t) - g_4(t)) + nn_2(x)g_2(t) + nn_3(x)g_3(t) + nn_4 g_4(t)$$

Collecting the transition functions, one obtains

$$\gamma nn_1(x) + g_2(t)(nn_2(x) - nn_1(x)) + g_2(t)(nn_3(x) - nn_1(x)) + g_4(t)(nn_4(x) - nn_1(x))$$
Now assume
\[ nn_i(x) = (1 + \gamma)nn_1(x), \quad i = 2, \ldots, 4 \] (41)
then the model equation can be written as
\[ \gamma(1 + g_2(t) + g_3(t) + g_4(t))nn_1(x) \] (42)
Note that the condition (41) assumes that for all four regimes, the neural network approximation differs only by a multiplicative constant. Assuming
\[ g_i(t) = c_i/(1 + \exp(-a_it - b_i)) \] (43)
then in full neural network notation, the formulation of Eq. (42) is
\[ nn(2, H, 1)(t) \times nn(I, H, I)(x) \] (44)
Assuming four times transitions \((H_i = 3)\), our neural network model is given by
\[ (t, x) \rightarrow nn(2, 3, 1)(t) \times nn(I, H, I)(x) \] (45)
with
\[ nn(2, 3, 1)(t) : t \rightarrow 1 + \sum_{i=1}^{3} \frac{c_i}{1 + e^{-a_it - b_i}} \] (46)
\[ nn(I, H, 1)(x) : x \rightarrow 1 + \sum_{i=1}^{H} \frac{g_i}{1 + e^{-a_ix - b_i}} \] (47)
The \(nn(2, 3, 1)(t)\) will be denoted as the \(t\)-network; as the size of this network is fixed, we will leave out from the notation the size-arguments, and write just \(nn(t)\). The function \(nn(I, H, 1)(x)\) will be called the \(x\)-network; like the \(t\)-network, the \(x\)-network is denoted as \(nn(x)\). However, in this case we shall explicitly mention the size of this network, especially the value of \(H\).
Hence, instead of Eq. (44), we write in short:
\[ nn(t) \times nn(x) \] (48)
The input variable \(x\) will be (again) the data series \(LHUR(-1)\); hence, the \(x\)-network has \(I = 2\), and the input data \(t\) will be a time-index series \(#\{1, 2, \ldots\\}\. As before, \(x\) data are scaled down to \([0.1, 0.9]\). However, for convenience, in graphs the data are scaled back to their original range.
The number of hidden layer cells $H$ in the $x$-network $nn(x)$ has to be defined. We report on the results for $H=3$ and $H=1$. Networks with more hidden layer cells do not give other results in the sense that the neural networks generate outputs similar to the one depicted in Fig. 12; additional hidden layer cells have a low incremental contribution.

Results on $H=3$.

Statistical values: $MUSSR = 0.002$, $R^2 = 0.96$.

In Fig. 12 the estimated- and actual data are shown.

The performance is much better (as expected) than the foregoing network approximation.

In order to find out how the time network performs and whether an elementary “Phillips curve” is found by this model, the time-series outputs of the $t$-network $nn(t)$ and $x$-network $nn(x)$ are graphed separately (see Fig. 13).*

It seems that the $x$-network has indeed found some basic Phillips-curve structure. However, the $t$-network shows only three levels while

*All data in the figures are on the same scale: all data are scaled back to the original range.
FIGURE 13  Actual data INFR12 (dots) compared to output of t-network $nn(t)$ (left) and $x$-network $nn(x)$ (case $H = 3$) (right).
from Fig. 11 four levels may be expected. So, one transition is missing in the t-network. Looking at the output of each hidden layer cell of the t-network \( nn(t) \) separately reveals the same: hidden cells 1 and 2 will generate three levels while hidden cells 3 is almost linear in time (see Fig. 14) (the almost flat zero line in the figures is the graph of actual inflation rate data \( \text{INFR}12 \)).

Examining the \( x \)-network \( nn(x) \) reveals that this network has a level transition itself; such a level transition is made up by two hidden cells with almost equal but reverse in sign output level. So, an \( x \)-network with only one hidden cell (\( H = 1 \)) is tried out. In that case, level transitions, if found by the combined network \( nn(t) \times nn(x) \), may only be found in the \( t \)-network.

Results on \( H = 1 \).
Statistical values: \( \text{MUSSR} = 0.0023, \quad R^2 = 0.93 \).
In Fig. 15 the estimated- and actual data \( \text{INFR}12 \) are shown.
The output of the \( t \)-network \( nn(t) \) and \( x \)-network \( nn(x) \) show the three transitions and in the case of the \( x \)-network, a smooth “Phillips curve” (see Fig. 16).
It seems that the $x$-network has indeed found (again) some basic Phillips-curve structure and now the time network $nn(t)$ indeed generates four levels. Approximately, those time transitions occur where the input to a hidden cell will be zero, or otherwise stated, in the point of symmetry of the logistic function. In the case of the $t$-network, this will be at $t_i = -b_i/a_i$ [see Eq. (46)]. In this case the transitions take place at

$$
\begin{align*}
  t_1 &= 171.93 \sim \text{April 1974} \\
  t_2 &= 300.82 \sim \text{February 1982} \\
  t_3 &= 425.13 \sim \text{June 1995}
\end{align*}
$$

The data $t_1$ corresponds more or less to the first oil crises, data $t_2$ corresponds to a business cycle slowdown in the U.S. economy while the third data $t_3$ coincides with the beginning of a period of low inflation and low unemployment. Looking at the output of each hidden layer cell of the $t$-network $nn(t)$ separately reveals the same (see Fig. 17). We repeat that our method is descriptive. For a statistical approach which enables one to compute confidence intervals around these dates, we refer to White (2000).

We end this section with a similar remark as made at the end of Section 3 on nonlinear trends in real exchange rates. That is, our empirical analysis on the Phillips curve indicates that the unemployment–inflation
FIGURE 16 Output of $p(t)$ (left) and $r(t|x)$ (case $H = 1$) (right) compared to actual data [WFR12].
trade-off is shifting over time. We have also presented an approximation of the dates of these transitions. It is of considerable interest to find an economic explanation of these time transitions. This is a topic which is, however, beyond our exploratory data analysis.

5 CONCLUSION
In this chapter we gave a brief exposition of neural networks and their flexibility in handling complex patterns in economic data. A descriptive method to prune the size of neural nets in order to avoid overfitting is summarized. It is shown how a neural network is used to recover the dynamic properties of a nonlinear system, in particular, its stability by making use of the Lyapunov exponent. A two-stage network has been introduced where the usual nonlinear model is combined with time transitions, which may be handled by neural networks. The empirical examples on nonlinear trends in real exchange rates and a time-varying Philips curve using U.S. data indicate the applicability of the proposed procedures. Further research is needed to allow for more than one output
and it is a challenge for neural network analysis to recover common nonlinear trends in multivariate nonlinear systems.

ACKNOWLEDGMENTS

The authors are indebted to Lennart Hoogerheide, Rutger van Oest, and Duan Ooms for helpful assistance.

REFERENCES


13

Real-Time Forecasting with Vector Autoregressions: Spurious Drift, Structural Change, and Intercept Correction

Ronald Bewley
The University of New South Wales, Sydney, Australia

1 INTRODUCTION

In its simplest form, a vector autoregression (VAR) is an unrestricted reduced-form model that expresses each variable as a linear function of a constant and the lags of that and each other variable in the system. Since each equation in a VAR has the same regressors, they can be estimated separately by OLS. However, even in moderately sized systems with, say, six variables and four lags of each, and a constant term, there are 25 parameters to be estimated in each equation so that over-parameterization has often been cited as the main cause of poor forecasting performance.

Since their introduction in the late 1970s, one strand of research has focused on reducing the parameter space by imposing stochastic restrictions, which imply that each variable follows a random walk with drift (Litterman, 1980, 1986; Doan et al., 1984). These restrictions are justified by appealing to the common observation that most economic time series are integrated to order one, I(1). The combination of these priors...
with data-determined parameter estimates produces the variant commonly
known as the BVAR, or Bayesian VAR, with the so-called Litterman
priors after its originator. However, the estimation procedure is more
properly known as mixed estimation (Theil and Goldberger, 1961) and it
can trivially be implemented using OLS with a data-augmentation dummy
variable procedure (Robertson and Tallman, 1999).

The original BVAR models have become commonplace in the
literature with their relative success usually being attributed to the general
reduction in the parameter space (e.g., Artis and Zhang, 1990; Funke,
1990; Holden and Broomhead, 1990; Shoesmith, 1995). In a dissenting
note, Bewley (2000a) argues that this improvement is more likely due to
the correction of the downward bias in the estimated unit roots (or near-
unit roots) that has been shown to increase with the number of variables
in the system (Abadir et al., 1999). More recently, Sims and Zha (1998)
have further developed the BVAR methodology with priors that have
more general applicability and an estimation procedure that is more in the
true Bayesian spirit.

A separate line of enquiry has focused on pretesting the time
series for unit roots and, where appropriate, testing for cointegration
in the spirit of Engle and Granger (1987) and Johansen (1988). While
these methods typically result in only a minimal reduction in the
parameter space, the imposition of unit roots have been shown to pro-
duce marked improvements in forecast accuracy over both an unre-
stricted VAR and a BVAR with Litterman priors (Bewley, 2000a, 2000b).
The main justification for this result is that forecasts of \( I(1) \) vari-
ables from vector error correction (VEC) models rapidly approach linear
time trends. Therefore, it is the precision of the estimate of the constant
term, rather than short-run dynamics, that is the key issue in forecasting
with \( I(1) \) time series (Clements and Hendry, 1995, 1998b; Bewley,
2000b).

Clements and Hendry’s (1995, 1998b) taxonomy of the sources of
forecast errors shows that, if it is assumed that the data-generating process
(dgp) is stable over time, and a consistent estimator is used, there are only
three important sources of forecast error in a VAR in levels with \( I(1) \) data:
the failure to impose unit roots; over-parameterization arising from an
unrestricted reduced form specification; and, importantly, the precision in
estimation of the drift parameters. In the standard VAR framework, the
constant terms are nonlinear functions of all of the drift parameters and
all of the (many) coefficients on the lagged variables, so that it is a non-
trivial matter to address the question of controlling the estimates of drift
in that framework. However, Bewley (2000b) proposed a new representa-
tion which allows a simple, possibly Bayesian, direct estimation of the drift
parameters in a VEC. Thus, in case where some variables contain drift while others do not, the so-called mixed drift case, substantial improvements in forecast accuracy can be made by imposing exact, or stochastic, restrictions on the drift parameters.

Since the estimation of drift is of such importance in VEC models, Bewley and Yang (2000) developed a number of statistics to test for structural breaks specifically in that parameter vector. While system tests previously existed (Andrews, 1993; Andrews and Ploberger, 1994), these new tests allow for a simple variable-by-variable test for parameter constancy that are more appropriate when only some of the variables are subject to structural breaks in drift.

The methods introduced in this chapter are designed to take a properly specified model that has been successfully subjected to appropriate diagnostic tests and use that model in a real-time setting. In practice, structural breaks are commonplace and potentially an extremely important contributor to poor forecasting performance. As a result, Hendry and Clements (1994), Clements and Hendry (1995, 1996, 1998a,b), and Hendry (1996, 1997) have suggested correcting the intercept at each forecast origin to realign the forecasts after a break has occurred within the sample. This correction is achieved by adding the most recent residual, or the average over the last few periods, to the one-step forecasts or to the forecasts at all lead times.

While Clements and Hendry advocate intercept corrections with VEC models, they argue that they have less merit in a VAR model in the differences (DVAR). Indeed, they consider a (non intercept corrected) DVAR, which results from a mis-specifying of a VEC model, as a viable alternative to intercept correcting a VEC model. In that sense, Clements and Hendry consider breaks in the long-run means of the equilibrium-(error-) correction terms as more important, and more likely, sources of forecast error than breaks in drift parameters. This view is supported by Eitrheim et al.’s (1990) experimentation with a macroeconometric model of the Norwegian Central Bank.

While there is clear merit in the Clements and Hendry prescription, the results in Bewley (2000b) suggest that a fruitful approach might be to allow the drift parameters, rather than the constant terms in a VEC, to change in the recent past in both a classical and a mixed-estimation sense. While these methods can be applied to either VEC or DVAR models, the emphasis in this chapter is on the latter. It is found that substantial improvements in forecast accuracy can be made by intercept correcting DVARs.

The proposed intercept correction methods are introduced in Section 2 for the general VEC model. The design of the simulation...
experiment is presented in Section 3 and the results are reported in Section 4. Conclusions are drawn in Section 4.

2 INTERCEPT CORRECTION

Consider an n×1 vector of I(1) times series, yt, that can be represented by a VAR with p lags, VAR(p):

\[ y_t = a + \sum_{i=1}^{p} A_i y_{t-i} + u_t \]  

(1)

If \( \sum A_i - I \) has rank \( r > 0 \), Eq. (1) can be written as VEC(p):

\[ \Delta y_t = a + \alpha \beta y_{t-1} + \sum_{i=1}^{p-1} B_i \Delta y_{t-i} + u_t \]  

(2)

where \( B_j = -\sum_{i=j+1}^{p} A_i \) and \( \sum_{i=1}^{p} A_i - I = \alpha \beta' \), with \( \alpha \) and \( \beta \) being \( n \times r \). If the time series are not cointegrated (\( r = 0 \)), then a VAR in \( \Delta y_t \) with \( p - 1 \) lags is appropriate, DVAR(p - 1).

Clements and Hendry (1998b) write Eq. (2) as

\[ [\Delta y_t - \delta] = \alpha \beta' y_{t-1} - \gamma + \sum_{i=1}^{p-1} B_i \Delta y_{t-i} - \delta + 1 \]  

(3)

where \( \gamma \) is the mean of the equilibrium-correction terms, \( \beta' y_{t-1} \), and \( \delta \) is the mean of \( \Delta y_t \), the drift parameters. The forecasts of Eq. (3) approach linear time trends with slope \( \delta \). If follows from a comparison of Eqs. (2) and (3) that the constant term, \( a \), in Eq. (2) is a nonlinear function of the drift, \( \delta \), and all of the \( B_j \):

\[ a = \left[ \left( I - \sum_{i=1}^{p-1} B_j \right) \delta - \alpha \gamma \right] \]  

(4)

Bewley (2000b) proposed an alternative representation of Eqs. (2) or (3) that allows direct estimation of the drift parameter vector, \( \delta \). This is achieved by applying the Bewley (1979) transformation to Eq. (3):

\[ \Delta y_t = \delta + \zeta \beta' y_{t-1} - \gamma + \sum_{i=1}^{p-2} C_i \Delta^2 y_{t-i} + v_t \]  

(5)

where the \( C_i \) are nonlinear functions of all of the \( B_j \) and \( \zeta \) is a nonlinear function of all the \( B_j \) and \( \alpha \). Since each Equation in (5) is exactly identified, two-stage least squares (2SLS), conditional on a superconsistent estimate of \( \beta \), and using a constant term, the equilibrium-correction terms
and the $p - 1$ lags of $\Delta y$, as instruments, is equivalent to indirect least squares, implying an exact relationship between the estimates of the three basic forms of a VEC: Eqs. (2), (3), and (5). Bewley (2000b) introduced Eq. (5) to enable some of the drift parameters to be set to zero with exact or stochastic restrictions in the mixed-drift situation. Equation (5) is referred to as the mixed-drift VAR (MDVAR).

When restrictions are placed on the MDVAR, the equations are no longer exactly identified and a system estimation approach, such as three-stage least squares (3SLS), or iterated 3SLS, is appropriate.

If it is assumed that the means of the equilibrium-correction terms are constant over time, but that all of the drift parameters have changed in a recent time period, an appropriately defined dummy variable can be added to each equation of (5) or, equivalently, Eqs. (2) or (3). This is one notion of intercept correction. If only a subset of the drift parameters has changed, it is not possible to impose linear restrictions on the dummy variable augmentation to Eq. (2) or (3). However, this is not the case in the MDVAR, and so possible gains might be had by exactly or stochastically restricting the dummy variable parameters in that framework.

On the other hand, if the long-run means of the equilibrium correction terms, $\zeta$, vary, it is not possible to distinguish between changes in the two long-run means, $\delta$ and $\zeta$ with a single dummy variable for each equation. Importantly, changes in any one element of $\zeta$ implies that all of the dummy variable coefficients are nonzero in either the VEC or MDVAR. In that sense, it is not clear that there is an advantage to pursuing the latter when equilibrium-correction terms are included and are possibly subject to structural breaks. The notion that the equilibrium-correction terms do not break, but the individual series do, has been termed co-breaking by Hendry and discussed in Clements and Hendry (1998b) and elsewhere.

Since it may be the case that only a subset of the time series has been subjected to a structural break, the mixed-break case, but that this is not known to the model builder, consideration of a Bayesian intercept-correction method, which introduces a dummy to the equations in (5), but shrinks the coefficients to zero, may prove fruitful. Such a procedure has two potential advantages. First, shrinking the intercept correction toward zero implies that the stochastically restricted MDVAR is less likely to overreact to noise. Second, those series without breaks would have dummy variable coefficients with large standard errors, resulting in restricted estimates that would be attracted toward zero more so than for series that did break. When a similarly defined Bayesian intercept correction is applied to a standard DVAR or VEC, the mixed-break case implies that all of
the dummy variable coefficients are typically nonzero so that this differential shrinking to zero cannot easily be exploited.

In order to implement this stochastic restriction on a DVAR or VEC, it is useful to consider the $i$th equation of a DVAR as a standard regression equation:

$$ w = X \pi + \varepsilon $$

(6)

using obvious notation (conditional on the equilibrium-correction terms having been defined) and assuming that the first column of $X$ is the dummy variable introduced for intercept correction. The stochastic restriction can be written as

$$ r = R \pi + \nu $$

(7)

where $r=0$ and $R$ is a row vector of zeroes except for the first element, which is unity.

It is necessary to prespecify the variance of $\nu$, which controls the so-called tightness of the prior; $\text{Var}(\nu) = \lambda^2$. Assuming $\text{Var}(\varepsilon) = \sigma^2 I$, stacking Eqs. (6) and (7) produces the augmented model:

$$ \begin{pmatrix} w \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \Pi + \begin{pmatrix} \varepsilon \\ \nu \end{pmatrix} $$

(8)

which might be estimated by GLS since $\sigma$ is not necessarily equal to $\lambda$. Alternatively, Eq. (7) can be multiplied through by $\hat{\sigma}/\lambda$ and this equation stacked with Eq. (6):

$$ \begin{pmatrix} w \\ r \end{pmatrix} = \begin{pmatrix} X \\ R^* \end{pmatrix} \Pi + \begin{pmatrix} \varepsilon \\ \nu^* \end{pmatrix} $$

(8')

where the first element of $R^*$ is $\hat{\sigma}/\lambda$ and the remaining elements are zero. The variance of $\varepsilon$ and $\nu^*$ are now approximately the same. Thus, Eq. (8) can be estimated by OLS; the so-called dummy variable augmentation procedure. In keeping with the BVAR literature, this estimate of $\sigma$ is taken from a $p-1$th order autoregression of $\Delta y_{i,t}$.

When exact or stochastic restrictions are applied to the MDVAR, certain modifications are necessary. The correlation of $X$ with $\varepsilon$ requires $X$ to be replaced by its least squares predictions $\hat{X}$ in typical 2SLS fashion. However, as $\lambda \to 0$, the resulting estimates do not approach those that would have been computed if the dummy variables has been excluded from the system, as would be the case if this were applied in the DVAR/VEC situation. This difference arises because the instruments used in creating $\hat{X}$ include the dummy variable which is then being “excluded” by the restrictions. The difference might be expected to be very small when the restriction is true but, when a significant break has occurred, sizeable
differences can emerge in parameter estimates and forecast accuracy. To rectify this problem, the first-stage regressions can be similarly augmented with the dummy variable procedure to shrink the first-stage dummy variable coefficients to zero. That is, the \( T+1 \)th observations are all zero except for that corresponding to the dummy variable that is set to \( \delta/\lambda \). In this way, the restricted estimates approach those from not including an intercept correction as \( \lambda \to 0 \) and the unrestricted estimates, including an intercept-correction term, as \( \lambda \to \infty \). While \( \delta \) is taken from autoregressions of \( \Delta y_t \), in the case of a DVAR/VEC, \( \delta \) is taken from the Bewley transformation of that autoregression in the MDVAR.

3 MONTE CARLO DESIGN

In a series of papers, Hendry, and Clements and Hendry, have argued that intercept corrections to VEC models perform on a par with DVARs without corrections and that corrections to the latter tend to reduce forecast performance due to an over-reaction to noise. Since the analysis of Section 2 suggests that there might be greater potential to make improvements in MDVARs than in DVARs or VEC models, the ensuing experiment has been designed to investigate the merits, if any, of intercept correcting an MDVAR when there is no cointegration. Thus, consider a two-equation mixed-break dgp with \( T \) observations and no cointegration:

\[
\Delta y_t = \delta_1 + \phi_1 (\Delta y_{t-1} - \delta_1) + \phi_2 (\Delta x_{t-1} - \delta_2 - \delta^* D_{t-1}) + \varepsilon_{1,t} \quad (9)
\]

\[
\Delta x_t = \delta_2 + \delta^* D_t + \phi_3 (\Delta x_{t-1} - \delta_2 - \delta^* D_{t-1}) + \varepsilon_{2,t} \quad (10)
\]

where \( \varepsilon_t \) is n.i.d \((0,1)\) and \( E(\varepsilon_1 \varepsilon_2) = 0 \). \( D_t \) is the dummy variable that introduces the structural break in the drift of \( x_t \) but not \( y_t \). Note that this dummy is lagged when it operates on \( \Delta x_{t-1} \) to align it with the timing of the break.

Equivalently, (9) and (10) can be expressed as

\[
\Delta y_t = \gamma_1 + \mu_1 D_{t-1} + \phi_1 \Delta y_{t-1} + \phi_2 \Delta x_{t-1} + \varepsilon_{1,t} \quad (9')
\]

\[
\Delta x_t = \gamma_2 + \delta^* D_t + \mu_2 D_{t-1} + \phi_3 \Delta x_{t-1} + \varepsilon_{2,t} \quad (10')
\]

where \( \gamma_1 = (1 - \phi_1) \delta_1 - \phi_2 \delta_2; \quad \gamma_2 = (1 - \phi_3) \delta_2; \quad \mu_1 = -\phi_2 \delta^*; \quad \mu_2 = -\phi_3 \delta^* \). Thus, both constants shift in a DVAR as shown in Eqs. (9’) and (10’), providing that \( \phi_2 \neq 0 \), but there is no observable shift in \( y_t \), because the shift in \( x_t \) is offset by the change in the drift term in Eq. (9). However, in a real-time forecasting situation, there is a delay of one period before the constant term in Eq. (9’) changes as, in the first period of the change,
there is a single unit element in $D_t$ corresponding to the last observation and, hence, in that special case $D_{t-1} = 0$.

It is assumed that $D_t = 0$ if $t < T - q$ and $D_t = 1$ if $t \geq q$. That is, the break in drift is permanent (and continues throughout the forecasting period). Moreover, $q$ is also assumed to be sufficiently small as to make the shift difficult to detect. Any previous changes in drift are assumed to have been modeled. Following the construction of the Andrews and Ploberger (1994) and Bewley and Yang (2000) Wald tests, breaks are assumed not to have occurred at the ends of the sample, say within 15% of the end of the sample. This serves as a guide to the value of $q$. Two values of $q$ are considered 0 and 4. Clearly, even if a significant shock were to be detected in the last observation, corresponding to $q = 0$, it would be impossible to test whether the shock was temporary or permanent. In two cases corresponding to $q = 0$ and 4, it is assumed in estimation that the location of the hypothesized break is known. In a third case, the assumed location, $q^*$, is set to 4 while the true value is $q = 0$. This serves two purposes. In reality, the location of a break would not be known and, second, it is consistent with Clements and Hendry’s approach of averaging recent errors so that overreaction to shocks is less prevalent.

Two basic dgos are considered. The first, referred to as the “diagonal dgp,” has $\phi_1 = \phi_3 = 0.5$ and $\phi_2 = 0$. This implies that the constant term in the first equation of the DVAR does not depend on the size of the break in the second, $\delta^*$; i.e., there is no “transmitted” structural break. In the second dgp, $\phi_1 = \phi_3 = 0.5$ so that the speed of adjustment to equilibrium is unchanged, but $\phi_2 = -0.8$, implying that both constant terms in the DVAR depend on $\delta^*$; i.e., there is a “transmitted” structural break in the first equation but not the first variable. Three values of $\delta^*$ are considered: 1, 2, and 4 and these compare to $\text{Var}(\Delta y_t) = \text{Var}(\Delta x_t) = 1/(1-0.5^2) = 4/3$. Thus, $\delta^* = 4$ is unlikely to be encountered in many applications but is included to assess the impact of the proposed methods in extrema.

Since none of the forecasting comparisons depend on the drift parameters, both were arbitrarily set equal to one. These values, together with the other parameters chosen for the dgp, produce time series that visually have the characteristics of many macroeconomic time series. That is, they exhibit strong drift, but deviations from a linear trend are pronounced. A stylized example for a single replication of the triangular dgp is given in Fig. 1 for $T = 85$ and $\delta^* = 0.1, 2, 4$ with a break occurring at observation 71. In the case of the diagonal dgp, the $y$ series has the same dgp as $x$ but without a break.

Each experiment is based on 5000 replications and $T = 75, 150$ with the forecast comparisons being based on the 1- and 10-step ahead mean
squared forecast error (msfe). In each case, the true lag length of one in first differences was chosen so as to abstract from any forecast inaccuracy due to selecting the wrong lag length. The DVAR was estimated with no correction and with an unrestricted intercept correction. The forecasts from these are necessarily the same as those from the comparable MDVARS. In addition, each form was estimated in its Bayesian form for 41 different values of $\lambda = (0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 15.0, 20.0, 25.0, 30.0, 35.0, 40.0, 50.0, 75.0, 100.0)$.

4 RESULTS

The 1-step and 10-step-ahead msfes for the DVAR with no intercept correction ($\lambda=0$) and with an unrestricted intercept correction ($\lambda \to \infty$), together with the Bayesian variant of the DVAR and MDVAR that produced the smallest msfe for that estimator and that equation, are presented in Tables 1 and 2 for the case $T=75$.

When there is no transmitted drift, as there is in the diagonal dgp, the first variable has no structural break and, therefore, it would be expected that the no intercept correction model would produce the smallest msfe. On the other hand, the second variable does exhibit a break,
<table>
<thead>
<tr>
<th>$\delta^*$</th>
<th>$q^*$</th>
<th>$q$</th>
<th>dgp</th>
<th>DVAR ($\lambda = 0$)</th>
<th>DVAR</th>
<th>MDVAR</th>
<th>DVAR ($\lambda = \infty$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0 diag 1.05</td>
<td>0.1</td>
<td>1.05</td>
<td>0.7</td>
<td>1.04</td>
<td>2.01</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 diag 1.09</td>
<td>0.1</td>
<td>1.09</td>
<td>0.8</td>
<td>1.07</td>
<td>2.05</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 diag 1.23</td>
<td>0.1</td>
<td>1.23</td>
<td>1.0</td>
<td>1.20</td>
<td>2.19</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4 diag 1.05</td>
<td>0.1</td>
<td>1.05</td>
<td>0.4</td>
<td>1.05</td>
<td>1.24</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4 diag 1.09</td>
<td>0.1</td>
<td>1.09</td>
<td>0.4</td>
<td>1.08</td>
<td>1.28</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4 diag 1.23</td>
<td>0.1</td>
<td>1.23</td>
<td>0.5</td>
<td>1.23</td>
<td>1.43</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4 diag 1.04</td>
<td>0.1</td>
<td>1.04</td>
<td>0.5</td>
<td>1.04</td>
<td>1.22</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4 diag 1.07</td>
<td>0.1</td>
<td>1.07</td>
<td>0.5</td>
<td>1.06</td>
<td>1.22</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4 diag 1.13</td>
<td>0.1</td>
<td>1.13</td>
<td>0.5</td>
<td>1.13</td>
<td>1.23</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0 tri 1.73</td>
<td>0.1</td>
<td>1.73</td>
<td>1.3</td>
<td>1.38</td>
<td>2.73</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0 tri 3.78</td>
<td>0.1</td>
<td>3.78</td>
<td>3.0</td>
<td>1.68</td>
<td>4.78</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 tri 11.88</td>
<td>0.0</td>
<td>11.88</td>
<td>3.5</td>
<td>3.10</td>
<td>12.90</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0 tri 1.73</td>
<td>0.1</td>
<td>1.73</td>
<td>1.3</td>
<td>1.38</td>
<td>2.73</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 tri 3.78</td>
<td>0.0</td>
<td>3.78</td>
<td>1.7</td>
<td>3.25</td>
<td>4.00</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 tri 11.88</td>
<td>0.0</td>
<td>11.88</td>
<td>2.0</td>
<td>9.68</td>
<td>12.16</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0 tri 1.58</td>
<td>1.5</td>
<td>1.24</td>
<td>2.5</td>
<td>1.16</td>
<td>1.25</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0 tri 2.77</td>
<td>$\infty$</td>
<td>1.33</td>
<td>3.5</td>
<td>1.22</td>
<td>1.33</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0 tri 4.50</td>
<td>$\infty$</td>
<td>4.31</td>
<td>1.57</td>
<td>1.57</td>
<td>1.57</td>
</tr>
<tr>
<td>10-step</td>
<td>1</td>
<td>0</td>
<td>0 diag 39.19</td>
<td>0.0</td>
<td>39.19</td>
<td>0.2</td>
<td>39.17</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 diag 39.73</td>
<td>0.0</td>
<td>39.73</td>
<td>0.3</td>
<td>39.70</td>
<td>405.19</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 diag 42.08</td>
<td>0.0</td>
<td>42.08</td>
<td>0.3</td>
<td>41.99</td>
<td>524.02</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4 diag 39.19</td>
<td>0.0</td>
<td>39.19</td>
<td>0.2</td>
<td>39.16</td>
<td>100.67</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4 diag 39.73</td>
<td>0.0</td>
<td>39.73</td>
<td>0.2</td>
<td>39.69</td>
<td>103.31</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4 diag 42.08</td>
<td>0.0</td>
<td>42.08</td>
<td>0.2</td>
<td>42.02</td>
<td>114.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4 diag 39.19</td>
<td>0.0</td>
<td>39.19</td>
<td>0.2</td>
<td>39.18</td>
<td>100.15</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4 diag 39.73</td>
<td>0.0</td>
<td>39.73</td>
<td>0.2</td>
<td>39.15</td>
<td>100.15</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4 diag 42.08</td>
<td>0.0</td>
<td>42.08</td>
<td>0.2</td>
<td>42.02</td>
<td>114.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0 tri 122.88</td>
<td>0.0</td>
<td>122.88</td>
<td>1.2</td>
<td>119.81</td>
<td>1516.56</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 tri 160.11</td>
<td>0.0</td>
<td>160.11</td>
<td>1.5</td>
<td>143.18</td>
<td>3338.25</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0 tri 311.47</td>
<td>0.0</td>
<td>311.47</td>
<td>1.5</td>
<td>251.77</td>
<td>10616.73</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4 tri 122.88</td>
<td>0.0</td>
<td>122.88</td>
<td>0.7</td>
<td>121.79</td>
<td>316.26</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4 tri 160.11</td>
<td>0.0</td>
<td>160.11</td>
<td>0.8</td>
<td>155.70</td>
<td>480.62</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4 tri 311.47</td>
<td>0.0</td>
<td>311.47</td>
<td>0.8</td>
<td>296.54</td>
<td>1148.62</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4 tri 122.88</td>
<td>0.0</td>
<td>122.88</td>
<td>1.1</td>
<td>112.95</td>
<td>281.68</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4 tri 153.00</td>
<td>0.0</td>
<td>153.00</td>
<td>1.3</td>
<td>121.95</td>
<td>336.38</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4 tri 268.94</td>
<td>0.0</td>
<td>268.94</td>
<td>1.6</td>
<td>192.50</td>
<td>497.74</td>
</tr>
</tbody>
</table>
### TABLE 2  Mean Squared Forecast Errors for Variable 2 (break in drift)

<table>
<thead>
<tr>
<th>$\delta^*$</th>
<th>$q^*$</th>
<th>$q$</th>
<th>dgp</th>
<th>DVAR $\lambda = 0$</th>
<th>MSE $\lambda = 0$</th>
<th>MDVAR $\lambda = 0$</th>
<th>DVAR $\lambda = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-step</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>1.35</td>
<td>0.6</td>
<td>1.22</td>
<td>1.6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>2.16</td>
<td>0.9</td>
<td>1.32</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>5.47</td>
<td>1.1</td>
<td>1.47</td>
<td>3.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>1.35</td>
<td>0.4</td>
<td>1.32</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>2.16 $\infty$</td>
<td>1.82</td>
<td>6.0</td>
<td>1.82</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>5.47 $\infty$</td>
<td>3.45 $\infty$</td>
<td>3.45</td>
<td>3.45</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>1.29</td>
<td>0.5</td>
<td>1.17</td>
<td>1.3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>1.78</td>
<td>0.8</td>
<td>1.21</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>2.66</td>
<td>1.2</td>
<td>1.24</td>
<td>4.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>1.34</td>
<td>0.6</td>
<td>1.21</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>2.13</td>
<td>0.9</td>
<td>1.32</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>5.36</td>
<td>1.1</td>
<td>1.49</td>
<td>3.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>1.34</td>
<td>0.4</td>
<td>1.31</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>2.13 $\infty$</td>
<td>1.78</td>
<td>5.0</td>
<td>1.78</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>5.36 $\infty$</td>
<td>3.34 $\infty$</td>
<td>3.34</td>
<td>3.34</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>1.28</td>
<td>0.5</td>
<td>1.16</td>
<td>1.4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>1.75</td>
<td>0.8</td>
<td>1.20</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>2.51</td>
<td>1.2</td>
<td>1.22</td>
<td>4.5</td>
</tr>
<tr>
<td>10-step</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>114.37</td>
<td>0.6</td>
<td>78.97</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>344.50</td>
<td>0.9</td>
<td>110.47</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>1265.87</td>
<td>1.1</td>
<td>152.36</td>
<td>3.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>114.37</td>
<td>0.3</td>
<td>106.20</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>344.50 $\infty$</td>
<td>235.24</td>
<td>7.0</td>
<td>234.90</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>1265.87 $\infty$</td>
<td>629.90 $\infty$</td>
<td>629.90</td>
<td>629.90</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>103.84</td>
<td>0.5</td>
<td>66.69</td>
<td>1.3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>288.39</td>
<td>0.8</td>
<td>78.56</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>949.37</td>
<td>1.1</td>
<td>86.82</td>
<td>5.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>115.13</td>
<td>0.6</td>
<td>78.56</td>
<td>1.6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>347.46</td>
<td>0.9</td>
<td>110.48</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>1277.64</td>
<td>1.1</td>
<td>159.79</td>
<td>3.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>115.13</td>
<td>0.4</td>
<td>106.45</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>347.46 $\infty$</td>
<td>238.43</td>
<td>7.0</td>
<td>238.13</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>1277.64 $\infty$</td>
<td>651.79 $\infty$</td>
<td>651.79</td>
<td>651.79</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>105.50</td>
<td>0.5</td>
<td>66.03</td>
<td>1.4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>298.05</td>
<td>0.8</td>
<td>77.46</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>922.26</td>
<td>1.2</td>
<td>85.30</td>
<td>5.0</td>
</tr>
</tbody>
</table>
but it does not follow that an unrestricted intercept correction is to be preferred. This is the essence of the Clements and Hendry view that DVARs should not be intercept corrected. However, as the size of the break increases, unrestricted correction would become increasingly preferable.

From Table 1 it can be noted that both the 1-step and 10-step forecasts with the diagonal dgp do indeed have lower msfes when \( \lambda = 0 \) than when \( \lambda \rightarrow \infty \), and that the estimated \( \lambda \) is either zero or very small in the case of the DVAR. However, very small gains in msfe are apparent in the MDVAR with most of the estimated values of \( \lambda \) being in the approximate range 0.2–0.5 for 1-step ahead forecasts. Small values of \( \lambda \) are also estimated at the longer lead time.

Turning to Table 2, it can be noted that the ranking of the forecast performance for the second variable, and the diagonal dgp for the cases of \( \lambda = 0 \) and \( \lambda \rightarrow \infty \), depends on the size of the break, \( \delta^* \), the number of periods for which the break has been in operation, \( q \), and the number of periods for which the dummy variable assumes a change, \( q^* \). Not surprisingly, it is better to intercept correct (in the unrestricted sense) when \( \delta^* \) is large, except when the break occurred in the last observation and residual averaging \( (q^* > 0) \) does not occur. Whether or not \( \lambda \) is estimated in the DVAR or the MDVAR form, the msfe is never lower in either extreme and, sometimes, the gains from the Bayesian procedure are very substantial and greater than that achieved by residual averaging.

In the case of the diagonal dgp and variable 2, it can be also noted from Table 2 that the estimated values of \( \lambda \) are typically larger than those in Table 1, reflecting the existence of the structural change and its accommodation with an intercept correction. Most values for the DVAR are in the range 0.5–1.1 while that range is approximately 1.5–5.0 for the MDVAR.

The results for the triangular dgp are quite different from those with a diagonal dgp. In the triangular case, both equations in the DVAR experience a structural break while only the second breaks in the MDVAR form. For variable 1, unrestricted intercept correction is preferable to no correction in the case \( q^* = q = 4 \), but the reverse is true for the longer lead time. In most cases the Bayesian DVAR performed similarly to the no-correction DVAR and substantial gains can be found in the MDVAR variant. The estimated values of \( \lambda \) are mostly in the range 0.7–3.0 for the Bayesian MDVAR.

Not surprisingly, some of the greatest gains in msfe can be found in the triangular dgp and variable 2, and the Bayesian MDVAR is far superior to the Bayesian DVAR in many of the cases. Indeed, the Bayesian MDVAR produces the smallest (or equivalent) msfe in all but one of the 72 cases reported in Tables 1 and 2.
Two problems arise from the comparisons apparent in Tables 1 and 2. First, the value of $\lambda$ is not restricted to be the same in the two equations. Unless the model builder has different information about the likelihood of a break in each variable, a common value of $\lambda$ should be used. Second, it is not possible to ascertain the sensitivity of the msfe for $\lambda$ in the neighborhood of the minimum.

The former problem can be addressed by combining the msfe for the two equations for a common value of $\lambda$. Naturally, the results depend on the weights attached to each series. In Table 3, the two msfes are simply added, reflecting the similar trends, scale, and variability of the two time series. In no case is the Bayesian DVAR, the unrestricted intercept correction, or the no-correction DVAR, preferred to the Bayesian MDVAR and, in some cases, the gains are very large indeed. The greatest gains are made when the break is very large, $\delta^* = 4$, and the dgp is triangular. However, even when $\delta^* = 1$, $q = q^* = 4$, and the dgp is triangular, the gain in msfe for 10-step ahead forecasts of the Bayesian MDVAR is 20% over not correcting and 50% over intercept correction.

The sensitivity of the results to changes in $\lambda$ is considered in Figs. 2–9 for a break of $\delta^* = 2$ and $q = q^* = 4$. In each case there are three pairs of lines corresponding to variable 1, with the suffix (1), variable 2, with the suffix (2), and the sum of two msfes, with the suffix (1 + 2). Within each pair, one line refers to the Bayesian DVAR and the other to the Bayesian MDVAR. A log scale is used for $\lambda$ to allow for additional detail at smaller values of $\lambda$.

It can be seen by comparing Figs. 2–9 that there is a pronounced minimum sum of msfes for the Bayesian MDVAR. In some cases, such as in Fig. 7, the scale of the graph masks the full extent of the gain for that model. These minima also occur for variable 2, but it is the combination of the flatness of the curve for variable 1 over a wide range of $\lambda$, and the minimum of the curve for variable 2, that make the overall gain so pronounced.

The Bayesian MDVAR produces a lower sum of msfe, particularly at the longer lead time, and for a reasonably wide range of $\lambda$. In each case, an arbitrarily selected value of 1 would have produced a forecasting model preferable to either an unrestricted correction or no correction. While a more extensive set of experiments is warranted, the results of this design are sufficiently encouraging to explore the use of the Bayesian intercept correction MDVAR in practice.

The results for $T=150$ are not presented, but are available from http://economics.web.unsw.edu.au/people/rbewley, owing to their similarity with those given here. In essence, the additional observations improve the precision of the estimates of the basic DVAR but there are still...
<table>
<thead>
<tr>
<th>δ*</th>
<th>q*</th>
<th>q</th>
<th>dp</th>
<th>DVAR (λ = 0)</th>
<th>DVAR</th>
<th>MDVAR</th>
<th>DVAR (λ = ∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>λ</td>
<td>mse</td>
<td>λ</td>
<td>mse</td>
</tr>
<tr>
<td>1-step</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>2.40</td>
<td>0.5</td>
<td>2.31</td>
<td>1.3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>3.25</td>
<td>0.8</td>
<td>2.55</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>6.70</td>
<td>1.1</td>
<td>2.97</td>
<td>3.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>2.40</td>
<td>0.2</td>
<td>2.38</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>3.25</td>
<td>0.7</td>
<td>3.04</td>
<td>1.8</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>6.70</td>
<td>∞</td>
<td>4.87</td>
<td>∞</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>2.33</td>
<td>0.4</td>
<td>2.25</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>2.85</td>
<td>0.7</td>
<td>2.35</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>3.79</td>
<td>1.1</td>
<td>2.43</td>
<td>4.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>3.07</td>
<td>0.5</td>
<td>2.97</td>
<td>1.7</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>5.91</td>
<td>0.8</td>
<td>5.18</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>17.23</td>
<td>1.1</td>
<td>13.53</td>
<td>3.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>3.07</td>
<td>0.3</td>
<td>3.06</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>5.91</td>
<td>0.7</td>
<td>5.68</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>17.23</td>
<td>∞</td>
<td>15.50</td>
<td>2.5</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>2.86</td>
<td>0.9</td>
<td>2.44</td>
<td>1.7</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>4.51</td>
<td>3.5</td>
<td>2.63</td>
<td>3.0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>7.01</td>
<td>∞</td>
<td>2.96</td>
<td>7.0</td>
</tr>
<tr>
<td>10-step</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>153.57</td>
<td>0.4</td>
<td>132.40</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>384.23</td>
<td>0.7</td>
<td>201.87</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>diag</td>
<td>1307.95</td>
<td>1.0</td>
<td>316.36</td>
<td>3.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>153.57</td>
<td>0.2</td>
<td>149.47</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>384.23</td>
<td>0.7</td>
<td>318.83</td>
<td>1.8</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>diag</td>
<td>1307.95</td>
<td>∞</td>
<td>744.23</td>
<td>∞</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>143.02</td>
<td>0.4</td>
<td>118.49</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>328.14</td>
<td>0.6</td>
<td>149.87</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>diag</td>
<td>892.38</td>
<td>1.0</td>
<td>174.36</td>
<td>4.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>238.01</td>
<td>0.2</td>
<td>234.99</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>507.57</td>
<td>0.3</td>
<td>486.34</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>tri</td>
<td>1589.10</td>
<td>0.4</td>
<td>1493.40</td>
<td>2.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>238.01</td>
<td>0.1</td>
<td>237.50</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>507.57</td>
<td>0.2</td>
<td>499.66</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>tri</td>
<td>1589.10</td>
<td>0.4</td>
<td>1523.64</td>
<td>2.0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>226.76</td>
<td>0.3</td>
<td>215.36</td>
<td>1.3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>451.04</td>
<td>0.5</td>
<td>336.04</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>tri</td>
<td>1191.20</td>
<td>0.9</td>
<td>544.45</td>
<td>3.5</td>
</tr>
</tbody>
</table>
FIGURE 2  One-step-ahead msfe for diagonal dgp and $\delta^* = 2$, $q^* = q = 0$.

FIGURE 3  Ten-step-ahead msfe for diagonal dgp and $\delta^* = 2$, $q^* = q = 0$. 
FIGURE 4 One-step-ahead msfe for diagonal dgp and $\delta^* = 2$, $q^* = q = 4$.

FIGURE 5 Ten-step-ahead msfe for diagonal dgp and $\delta^* = 2$, $q^* = q = 4$. 
FIGURE 6 One-step-ahead msfe for triangular dgp and $\delta^* = 2$, $q^* = q = 0$.

FIGURE 7 Ten-step-ahead msfe for triangular dgp and $\delta^* = 2$, $q^* = q = 0$. 
FIGURE 8 One-step-ahead msfe for triangular dgp and $\delta^* = 2$, $q^* = q = 4$.

FIGURE 9 Ten-step-ahead msfe for triangular dgp and $\delta^* = 2$, $q^* = q = 4$. 
only one \( q = 0 \) or five \( q = 4 \) observations in the new regime after the break. In that sense, the forecast relativities of the model do not depend on sample size.

5 CONCLUSIONS

A new strategy has been proposed for estimating and testing VAR forecasting models, including the possibility of structural change, in a series of papers: Bewley (2000a, 2000b) and Bewley and Yang (2000), and the present chapter. The essence of the approach is to focus attention on the drift parameters. Spurious drift, arising from poorly determined drift estimates of variables that would not be expected to trend, cause problems of forecast credibility and, while relevant restrictions could be placed on a standard VEC or DVAR model, the nonlinearity of the restrictions make such estimation cumbersome and prone to optimization problems. The proposed framework for estimating VARs—the mixed-drift VAR (MDVAR)—facilitates this estimation and has been shown to permit more accurate and more credible forecasts to be produced (Bewley, 2000b), using either exact or stochastic restrictions.

The computational time to estimate the proposed parameterization is trivial owing to there being an analytical solution. However, there is much scope for experimenting with different priors when stochastic restrictions are being contemplated, but even this is quite feasible. For example, 5000 replications, 41 sets of priors, 75–150 observations, and two equations took only 13 min with a Pentium II processor and a program written in FORTRAN.

In a natural extension to Bewley (2000b), Bewley and Yang (2000) proposed a set of statistics to test for structural change specifically in the drift parameters. The MDVAR framework simplifies the allocation of any structural change in the system to attribute it to possibly a subset of the variables. The present chapter considers possible structural change in the most recent observations, and an automatic Bayesian procedure for adapting to structural change in drift, without overreacting to noise, is proposed. It has been found that, contrary to previous research (Clements and Hendry), significant gains in mean squared forecast error can be made by intercept correcting a VAR in differences. Indeed, when there is a significant degree of causality in the system, and structural change is confined to a subset of the variables, substantial gains in forecast accuracy can occur with the new estimator.

While the usual caveat of drawing conclusions from Monte Carlo experiments is at least as applicable here as elsewhere, the results are
sufficiently encouraging to warrant further investigation. In particular, work is progressing on extending these results to discriminating between structural breaks in drift and in the long-run means of equilibrium-correction terms.

ACKNOWLEDGMENTS

This work was funded by an Australian Research Council grant (A10007132). This chapter has benefited greatly from many useful discussions with, and comments from, David Hendry, whom I thank. The procedures developed in this chapter were inspired by “Hans” Theil’s repeated least squares and mixed estimation innovations. Sadly, Hans, my dear friend, mentor, and sometime coauthor passed away while I was writing this chapter. I would like to acknowledge Hans’ support and encouragement over the years. I have also benefited from presentations at the UNSW time series group weekly workshops and the excellent research assistance provided by Glenn Anderson.

REFERENCES


Econometric Modeling Based on Pattern Recognition via the Fuzzy C-Means Clustering Algorithm

David E. A. Giles
University of Victoria, Victoria, British Columbia, Canada

Robert Draeseke
Treaty Negotiations, Ministry of Attorney General, Government of British Columbia, Victoria, British Columbia, Canada

1 INTRODUCTION

The specification and estimation of any econometric relationship poses significant challenges. This is especially true when it comes to the choice of functional form, as the latter is not always suggested or prescribed by the underlying economic theory. Any mis-specification of the functional form of an econometric model can have serious consequences for statistical inference—e.g., the parameter estimates may be inconsistent. In response to the rigidities associated with explicit and rigid parametric functional relationships, and to avoid the consequences of their mis-specification, a more flexible nonparametric approach to their formulation and estimation is often very attractive (e.g., Pagan and Ullah, 1999).

Such nonparametric estimation does, however, have its drawbacks. Most notably, in order to perform well, multivariate kernel regression requires a reasonably large sample of data, and it founders when the
number of explanatory variables grows. With regard to the latter point there are really two problems. First, there is evidence that the small-sample properties of the multivariate kernel estimator can be quite unsatisfactory (e.g., Silverman, 1986). Second, the rate of convergence of the kernel estimator decreases as the dimension of the model grows—this is the so-called “curse of dimensionality.” Recently, Coppejans (2000) has dealt with this problem by using cubic B-spline estimation in conjunction with earlier results from Kolmogorov (1957) and Lorentz (1966) that enable a function of several variables to be represented as superpositions of functions of a single variable.

There is considerable appeal in the prospect of finding approaches to the formulation and estimation of econometric relationships that are highly flexible in terms of their functional form, make minimal parametric assumptions, perform well with either small or large data sets, and are computationally feasible even in the face of a large number of explanatory variables. In this chapter we use some of the tools of fuzzy set theory and fuzzy logic in the pursuit of just such an objective.

These tools have been applied widely in many disciplines since the seminal contributions of Zadeh (1965, 1987) and his colleagues. These applications are numerous in such areas as computer science, systems analysis, and electrical and electronic engineering. The construction and application of “expert systems” is widespread in domestic appliances, motor vehicles, and commercial machinery. The use of fuzzy sets and fuzzy logic in the social sciences appears to have been limited mainly to psychology, with very little application to the analysis of economic problems and economic data. There are a few such examples in the field of social choice—e.g., Dasgupta and Deb (1996), Richardson (1998), and Sengupta (1999). In the area of econometrics, however, there have been surprisingly few applications of fuzzy set/logic techniques. Indeed, we are aware of only four such examples.

Josef et. al. (1992) used this broad approach in the context of modeling with panel data. Lindström (1998) used fuzzy logic in a rather different way to model fixed investment in Sweden on the basis of the level and variability in the real interest rate. More specifically, he generated an index for aggregate investment using fuzzy sets and operators to combine the imprecise information associated with the interest rate. He found that this approach provided a simple method of modeling inherent nonlinearities in a flexible (nonparametric) manner. It should be noted that Lindström’s analysis involved no statistical inference, as such. This same technique was used by Draeseke and Giles (1999, 2002) to generate an index of the New Zealand underground economy, and it has now been programmed into the SHAZAM (2000) econometrics package. The advantage of using this
particular technique in that context was that the variable of interest is intrinsically unobservable.

Shepherd and Shi (1998) used fuzzy set/logic theory in a rather different manner in their analysis of U.S. wages and prices. Their modeling technique involved clustering the data into fuzzy sets, estimating the relationship of interest over each set, and then combining each submodel into a single overall model by using the membership functions associated with the clustered data together with the Takagi and Sugeno (1985) approach to fuzzy systems. Their analysis also highlighted the ability of such fuzzy analysis to detect and measure nonlinear relationships in a very flexible way, even though the underlying submodels may be linear. Indeed, this is a specific and attractive feature of the Takagi–Sugeno analysis. In this chapter we draw on the approach of Shepherd and Shi and adapt it to model a number of interesting relationships. We discuss some of the issues that arise when adopting this semiparametric approach in the specification and estimation of an econometric model, and by way of specific practical applications we illustrate some of its merits relative to standard parametric regression and conventional (kernel) nonparametric regression.

The use of fuzzy sets and fuzzy logic is just one of several possible approaches within the general area of expert systems and artificial intelligence. In the context that is of interest to us here, an obvious competitive approach is that based on (artificial) neural networks (e.g., White, 1992; White and Gallant, 1992; Bishop, 1995). Neural networks also facilitate the modeling of nonlinear relationships whose underlying form is not parametrically constrained at the outset. However, our adaptation of the Takagi–Sugeno/Shepherd–Shi (TSSS) methodology has an important and appealing advantage over the use of neural networks. Namely, the basic relationships that are estimated have a direct economic interpretation, and they can be related to the underlying economic theory in a useful way. The same is not generally true in the case of neural networks. Moreover, and as we shall see in more detail below, the TSSS methodology is computationally efficient, especially in as much as it requires only one “pass” through the data. In contrast, neural networks usually have to be “trained,” and this involves more protracted computations.

In Section 2 we outline the background concepts in fuzzy set/logic theory that we will be using. A crucial component of our analysis is the implementation of the so-called “fuzzy c-means” algorithm to partition the data set in a flexible manner, and this algorithm is considered in some detail in Section 3. In Section 4 we describe our variant of the TSSS modeling procedure, and Section 5 presents the results associated with several illustrative empirical applications of this fuzzy technology. Some concluding remarks are given in Section 6.
2 FUZZY SETS AND FUZZY LOGIC

As was noted above, fuzzy logic relates to the notion of fuzzy sets, the theoretical basis for which is usually attributed to Zadeh (1965). Under regular set theory, elements either belong to some particular set or they do not. Another way of expressing this is to say that the "degree of membership" of a particular element with respect to a particular set is either unity or zero. The boundaries of the sets are hard, or "crisp." In contrast to this, in the case of fuzzy sets, the degree of membership may be any value on the continuum between zero and unity, and a particular element may be associated with more than one set. Generally, this association involves different degrees of membership with each of the fuzzy sets. Just as this makes the boundaries of the sets fuzzy, it makes the location of the centroid of the set fuzzy as well.

To consider an illustrative situation relating to a single economic variable, suppose we wish to distinguish between situations of excess supply and those of excess demand in relation to the price of some good. In traditional set theory we would have a situation such as

\[ S_s = \{ p : p > p^* \} \]
\[ S_d = \{ p : p < p^* \} \]
\[ S_e = \{ p^* \} \]

as the crisp sets representing those prices associated with excess supply, excess demand, and the equilibrium price \( p^* \), respectively. Any particular price, say $5, would definitely be in one and only one of these sets. That is, if its degree of membership with \( S_i \) is denoted \( u_i \), then \( u_j = 0 \) for all \( k \neq j \); for \( k, j = s, d, e \). In contrast to this, in a fuzzy set framework the sets \( S_s, S_d, \) and \( S_e \) would not have sharp boundaries and a particular price (such as $5) would be associated to some degree or other with each of these sets. For instance, its degrees of membership might be \( u_s = 0.2, u_d = 0.3, u_e = 0.5 \). Note that although the \( u_i \)'s sum to unity in this example, in general they need not, and they should not be equated with the "probabilities" that the price of $5 lies in each set.

It should also be noted that in the example above, the concepts involved are defined in a crisp manner: "excess supply," "excess demand," and "equilibrium." More generally, fuzzy set theory is just as capable of handling vague linguistic concepts, such as "a rather high price," or "a very low demand." So, we could broaden the above example to allow for fuzzy sets involving prices associated with "a very high excess supply," "a moderate excess supply," "a small excess supply," and so on. Again, the boundaries of these sets would be fuzzy, and degrees of membership would map prices to the sets.
The application of the inductive premise to fuzzy concepts poses some difficulties—not all of the usual laws of set theory are satisfied. In particular, the “law of the excluded middle” is violated, so a different group of set operators must be adopted. So, for example, the “union” operator is replaced by the “max” operator, “intersection” is replaced by “min,” and “complement” is replaced by subtraction from unity. Under these “fuzzy” operators the commutative, associative, distributive, idempotency, absorption, excluded middle, involution, and De Morgan’s Laws are all satisfied in the context of fuzzy sets. For example, if the universal set is $U = \{a, b, c, d\}$, let the fuzzy sets $S_1$ and $S_2$ be defined as $S_1 = \{0.4/a, 0.5/b, 1/d\}$ and $S_2 = \{0.1/a, 0.2/b, 0.5/c, 0.9/d\}$. Here, the numbers are the “degrees of membership,” which map the elements of $U$ to $S_1$ and $S_2$. Then, $A \cup B = \{0.4/a, 0.5/b, 0.5/c, 1/d\}$ and $A \cap B = \{0.1/a, 0.2/b, 0/c, 0.9/d\}$, $S_1' = \{0.6/a, 0.5/b, 1/c, 0/d\}$, and $S_2' = \{0.9/a, 0.8/b, 0.5/c, 0.1/d\}$.

Fuzzy sets can be used in conjunction with logical operators in a manner that will enable us to construct models in a flexible way. More specifically, we can use the degrees of membership that associate the values of one or more input variables with different fuzzy sets, together with logical “IF,” “THEN,” “AND” operations, to derive membership values that associate an output variable with one or more fuzzy sets. For example, in the case of the demand for a particular good, we might have a fuzzy rule of the form:

“IF the own-price of the good is quite low, AND the price of the only close substitute good is rather high, THEN the demand for the good in question will be fairly high.”

Or, to anticipate one of our empirical examples later in this chapter.

“IF the interest rate is fairly high, AND income (output) is relatively low, THEN the demand for money will be quite low.”

Of course, in general, there will be more than one fuzzy rule, and the output variable may have potential membership in more than one fuzzy set:

Rule 1: “IF the interest rate is fairly high, AND income (output) is relatively low, THEN the demand for money will be quite low.”

Rule 2: “IF the interest rate is average, AND income (output) is relatively high, THEN the demand for money will be moderately high.”

In such situations, a fuzzy outcome for the output variable could be inferred by taking account of the relevant membership values, and the
MAX/MIN operators associated with fuzzy sets. However, a fuzzy outcome is not usually adequate. For instance, in the last example, there is only limited interest in knowing that the fuzzy model predicts that in a certain period the demand for money will be “moderately high,” or “quite low.” It would be much more helpful to know, e.g., that the model predicted a demand of $10 million. In other words, we need a way of “defuzzifying” the predictions of the model.

The Takagi and Sugeno (1985) and Sugeno and Kang (1985) approach to dealing with this issue involves modifying the above methodology to one that involves rules of the form:

Rule 1: “IF the interest rate \( r \) is fairly high, AND income \( Y \) is relatively low, THEN the demand for money is \( M = f_1(r, Y) = M_1 \).”

Rule 2: “IF the interest rate \( r \) is average, AND income \( Y \) is relatively high, THEN the demand for money is \( M = f_2(r, Y) = M_2 \).”

Here, \( f_1 \) and \( f_2 \) are crisp, and generally parametric, functions that yield numerical values for \( M \). In their simplest form these functions would be linear relationships. The various values \( (M_1, M_2 \text{ etc.}) \) emerging from these rules can then be combined by taking a weighted average, based on the degrees of membership associated with the input variables and the fuzzy input sets. This is described in more detail in Section 4. First, however, we need to give further consideration to the definition and construction of the fuzzy input sets, and this involves grouping or “clustering” the input data appropriately.

3 THE FUZZY C-MEANS ALGORITHM

3.1 Overview of the Algorithm

In our modeling analysis, which is described in detail in Section 4, we need to determine the partitioning of the sample data for each explanatory (input) variable into a number of clusters. These clusters have “fuzzy” boundaries, in the sense that each data value belongs to each cluster to some degree or other. Membership is not certain, or “crisp.” Having decided on the number of such clusters to be used, some procedure is then needed to locate their midpoints (or more generally, their centroids) and to determine the associated membership functions and degrees of membership for the data points. To this end, Shepherd and Shi (1998) used a variant of the “fuzzy c-means” (FCM) algorithm. (The latter is sometimes termed the fuzzy k-means algorithm in the literature.) The FCM algorithm is really
a generalization of the “hard” c-means algorithm. It appears to date from Ruspini (1970), although some of the underlying concepts were explored by MacQueen (1967). The FCM algorithm is closely associated with such early contributors as Bezdek (1973) and Dunn (1973, 1974, 1977), and is widely used in such fields as pattern recognition, for instance.

The algorithm provides a method of dividing up the \( n \) data points into \( c \) fuzzy clusters (where \( c < n \)), while simultaneously determining the locations of these clusters in the appropriate space. The data may be multidimensional, and the metric that forms the basis for the usual FCM is “squared error distance.” The underlying mathematical basis for this procedure is as follows. Let \( x_k \) be the \( k \)th (possibly vector) data point \((k = 1, 2, \ldots, n)\). Let \( v_i \) be the center of the \( i \)th (fuzzy) cluster \((i = 1, 2, \ldots, c)\), let \( d_{ik} = \|x_k - v_i\| \) be the distance between \( x_k \) and \( v_i \), and let \( u_{ik} \) be the “degree of membership” of data point \( k \) in cluster \( i \), where

\[
\sum_{i=1}^{c} (u_{ik}) = 1
\]

The objective is to partition the data points into the \( c \) clusters and simultaneously locate those clusters and determine the associated “degrees of membership,” so as to minimize the functional:

\[
J(U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (d_{ik})^2
\]

There is no prescribed manner for choosing the exponent parameter, \( m \), which must satisfy \( 1 < m < \infty \). In practice, \( m = 2 \) is a common choice. In the case of crisp (hard) memberships, \( m = 1 \).

In broad terms, the FCM algorithm involves the following steps:

1. Select the initial location for the cluster centers.
2. Generate a (new) partition of the data by assigning each data point to its closest cluster center.
3. Calculate new cluster centers as the centroids of the clusters.
4. If the cluster partition is stable then stop. Otherwise go to step 2 above.

In the case of fuzzy memberships, the Lagrange multiplier technique generates the following expression for the membership values to be used at step 2 above:

\[
u_{ik} = 1/\left\{ \sum_{j=1}^{n} [(d_{jk})^2/(d_{ik})^2]^{1/(m-1)} \right\}
\]
Notice that a singularity will arise if \( d_{jk} = 0 \) in the above expression. This occurs if, at any point, a cluster center exactly coincides with a data point. This can be avoided at the start of the algorithm and generally will not arise subsequently in practice due to machine precision. If the memberships of data points to clusters are "crisp," then

\[
\begin{align*}
    u_{ik} &= 0; \quad \forall i \neq j \\
    u_{jk} &= 1; \quad \text{j.s.t. } d_{jk} = \min\{d_{ik}, i = 1, 2, \ldots, c\}
\end{align*}
\]

The updating of the cluster centers at step 3 above is obtained via the expression:

\[
v_i = \left[ \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \right]; \quad i = 1, 2, \ldots, c
\]

The fixed-point nature of this problem ensures the existence of a solution. See Bezdek (1981, Chap. 3) for complete and more formal mathematical details. As will be apparent, the FCM algorithm is simple to program. We have chosen to do this using standard commands in the SHAZAM (2000) econometrics package—this makes it convenient to compare our results with those from the other standard regression modeling techniques that are available in that package.

### 3.2 Some Computational Issues

Among the issues that arise in the application of the FCM are the following. First, a value for the number of clusters and the initial values for their centres have to be provided to start the algorithm. It appears that the results produced by the FCM algorithm can be sensitive to the choice of these start-up conditions. Second, being based on a "squared error" measure of distance, the algorithm can be sensitive to noise or outliers in the data.

In relation to the first of these issues, several proposals have been made. Among these are the following. Linde et al. (1980) proposed the binary splitting (BS) technique to initialise the clusters. The disadvantage of the BS technique is that it requires at least 10 or 20 data points in each cluster, and in practice this may not be attainable. Huang and Harris (1993) extended this approach to the "direct search binary splitting" technique. Tou and Gonzales (1974) proposed the "simple cluster-seeking" (SCS) method, which deals not only with the initial value issue, but also provides the clusters themselves. However, the SCS technique is not invariant to the order in which the data points are considered, and it also depends on certain “threshold” settings.
Yager and Filev (1992) introduced the so-called “mountain method” for deciding on the initial values. Their approach is apparently simple, but it becomes computationally burdensome in the context of multidimensional data. Consequently, Chiu (1994) proposed a modification of their approach, called “subtractive clustering,” that does not suffer the same problem. This is the approach used by Shepherd and Shi (1998). Babu and Murty (1993) used genetic algorithms to initialize the clusters, and Katsavounidis et al. (1994) proposed a further method (KKZ). More recently, Al-Daoud and Roberts (1996) proposed two initialization methods that appear to be especially useful in the context of very large data sets, and which out-performed both the KKZ and SCS techniques with their test data.

In relation to the second of the above issues (noise and outliers in the data), there have again been a number of contributions. Among these are the following. Following on from earlier (computationally expensive) contributions by Weiss (1988) and Jolion and Rosenfeld (1989), Davé (1991) modified the FCM algorithm (still retaining a “squared error” distance concept) so that noisy data points are effectively allocated to a “noise cluster,” rather than being allocated to other clusters, and so “contaminating” the latter. More recently, Keller (2000) suggested another related modification involving the use of a modified objective function with a weighting factor added for each data point—the objective being to assign a kind of influence factor to the single data points. Some authors have taken up the usual techniques used in the statistical literature to deal with the sensitivity of least squares to the problem of outliers. That is to say, methods of “robust estimation” have been incorporated into the FCM algorithm, and these have included “M-estimation,” “trimmed least squares,” and “least absolute deviations” alternatives to the least squares component of the FCM algorithm. A recent example of this approach is that of Frigui and Krishnapuram (1996), who integrate M-estimation into the FCM algorithm. They provide a computationally attractive generalization of the FCM algorithm that deals with both the identification of the number of clusters and the allocation of data points to these clusters simultaneously, and which is robust to noise and outliers.

4 FUZZY ECONOMETRIC MODELING

In this section we describe the TSSS methodology that we subsequently apply to a number of econometric model estimation problems. The discussion here is quite expository, and we begin by outlining the analysis in the very simple case where there is a single input variable (other than,
perhaps, a constant intercept). So, the fuzzy relationship is of the form:

\[ y = f(x) + \varepsilon \]

where the functional relationship will typically involve unknown parameters, and \( \varepsilon \) is a random disturbance term. While there is no need to make any distributional assumptions about the latter, if this is done then these can be taken into account in the subsequent analysis. If the disturbance has a zero mean, the fuzzy function represents the conditional mean of the output variable, \( y \). To this extent, the framework is the same as that which is adopted in nonparametric kernel regression.

The identification and estimation of the fuzzy model then proceeds according to the following steps:

**Step 1:** Partition the sample observations for \( x \) into \( c \) fuzzy clusters, using the FCM algorithm. This generates the membership values for each \( x \) value with respect to each cluster, and implicitly it also defines a corresponding partition of the data for \( y \).

**Step 2:** Using the data for each fuzzy cluster separately, fit the models:

\[ y_{ij} = f(x_{ij}) + \varepsilon_{ij}; \quad j = 1, \ldots, n_i; \quad i = 1, \ldots, c \]

In particular, if the chosen estimation procedure is least squares, then

\[ y_{ij} = \beta_{i0} + \beta_{i1}x_{ij} + \varepsilon_{ij}; \quad j = 1, \ldots, n_i; \quad i = 1, \ldots, c \]

**Step 3:** Model and predict the conditional mean of \( y \) using

\[ \hat{y}_k = \left[ \sum_{i=1}^{c} (\beta_{i0} + \beta_{i1}x_k)u_{ik} \right] / \left[ \sum_{i=1}^{c} u_{ik} \right]; \quad k = 1, \ldots, n \]

where \( u_{ik} \) is the degree of membership of the \( k \)th value of \( x \) in the \( i \)th fuzzy cluster, and \( \beta_{im} \) is the least squares estimator of \( \beta_{im}(m=0,1) \) obtained using the \( i \)th fuzzy partition of the sample.

**Step 4:** Calculate the predicted “input–output relationship” i.e., the derivative) between \( x \) and the conditional mean of \( y \):

\[ \frac{\partial \hat{y}_k}{\partial x_k} = \left[ \sum_{i=1}^{c} (b_{i1}u_{ik}) \right] / \left[ \sum_{i=1}^{c} u_{ik} \right]; \quad k = 1, \ldots, n \]

So, the fuzzy predictor of the conditional mean of \( y \) is a weighted average of the linear predictors based on the fuzzy partitioning of the explanatory
data, with the weights (membership values) varying continuously through the sample. This latter feature enables nonlinearities to be modeled effectively. In addition, it can be seen that the separate modeling over each fuzzy cluster involves the use of fuzzy logic of the form “IF the input data are likely to lie in this region, THEN this is likely to be the predictor of the output variable,” etc. The derivative of the conditional mean with respect to the input variable also has this weighted average structure and the same potential for nonlinearity.

Note that this modeling strategy is essentially a semiparametric one. The parametric assumptions could be relaxed further by using kernel estimation to fit each of the cluster submodels at Step 2 above, in which case the estimated derivatives (rather than coefficients) would be weighted at Steps 3 and 4, instead of the parameter estimates. However, some limited experimentation with this variation of the modeling methodology, in the context of the empirical applications described in Section 5, yielded results that were inferior to those based on least squares.

Of course, in general we will be concerned with models that have more than one input (explanatory) variable:

\[ y = f(x_1, x_2, \ldots, x_p) + \varepsilon \]

In such cases the steps in the above fuzzy modeling procedure are extended as follows, assuming a linear least squares basis for the analysis for expository purposes:

Step 1': Separately partition the \( n \) sample observations for each \( x_r \) into \( c_r \) fuzzy clusters (where \( r = 1, 2, \ldots, p \)), using the FCM algorithm. This generates the membership values for each observation on each \( x \) variable with respect to each cluster.

Step 2': Consider all \( c \) possible combinations of the fuzzy clusters associated with the \( p \) input variables, where

\[ c = \prod_{r=1}^{p} c_r \]

and discard any for which the intersections involve negative degrees of freedom \( (n_r < p) \). Let the number of remaining cluster combinations be \( c' \).

Step 3': Using the data for each of these \( c' \) fuzzy clusters separately, fit the models:

\[ y_{ij} = \beta_{0i} + \beta_{1i}x_{1ij} + \beta_{2i}x_{2ij} + \cdots + \beta_{pi}x_{pij} + \varepsilon_{ij}; \]

\[ j = 1, \ldots, n_i; \quad i = 1, \ldots, c' \]
Step 4': Model and predict the conditional mean of $y$ by using

$$
\hat{y}_k = \left[ \sum_{i=1}^{c'} (b_{iw} + b_{ik}x_{1k} + \cdots + b_{ip}x_{pk})w_{ik} \right] / \left[ \sum_{i=1}^{c'} w_{ik} \right],
$$

$k = 1, \ldots, n$

where

$$
w_{ik} = \prod_{r=1}^{p} \delta_{ij}u_{ijk}; \quad i = 1, \ldots, c'
$$

Here, $\delta_{ij}$ is a "selector" that chooses the membership value for the $i$th fuzzy cluster (for the $r$th input variable) if that cluster is associated with the $i$th cluster combination ($i=1, 2, \ldots, c$), and $b_{im}$ is the least squares estimator of $\beta_{im}$ obtained using the $i$th fuzzy partition of the sample.

Step 5': Calculate the predicted "input–output relationships" i.e., the derivatives) between the conditional mean of $y$ and each input variable, $x_r$:

$$
\left( \frac{\partial \hat{y}_k}{\partial x_{rk}} \right) = \left[ \sum_{i=1}^{c'} (b_{ir}w_{rik}) \right] / \left[ \sum_{i=1}^{c'} w_{ik} \right],
$$

$r = 1, \ldots, p; \quad k = 1, \ldots, n$

Comparing these steps with those in the case of a single input variable, it is clear that the computational burden associated with the fuzzy modeling increases at the same rate as in the case of multiple linear regression as additional explanatory variables are added to the model. Under very mild conditions on the input data and the random error term, fitting the submodels over each fuzzy cluster yields a weakly consistent predictor of the conditional mean of the output variable. The partitioning of the sample into fuzzy clusters, and the determination of the associated membership functions, involves using only the explanatory variable data in a nonstochastic manner. If the explanatory variables are exogenous then so will be the membership values that are used to construct the weighted averages of the least squares predictors in Step 4 above. Then, the fuzzy predictor of the conditional mean of $y$ at Step 4 will be weakly consistent.

5 SOME APPLICATIONS

We have applied the fuzzy econometric modeling described above to a number of illustrative estimation and smoothing problems, and the
associated results are described in this section. As was noted in the context of the FCM algorithm, the computation for these applications was undertaken by writing command code for the SHAZAM (2000) econometrics package. We have chosen a number of simple examples that involve both cross-section and time-series data, and which involve various degrees of model complexity in terms of the number of explanatory variables involved.

5.1 Modeling the Earnings–Age Profile

In our first application the relationship of interest is one that explains the logarithm of earnings as a function of age, the latter being a proxy for years of work experience. Typically, in the labor economics literature, this relationship has been modeled by using standard parametric regression, typically with both age and its square being included as regressors. This quadratic relationship allows for the fact that earnings would be expected to increase with age through much of the working life, but that the rate of increase declines with age (e.g., Heckman and Polachek, 1974; Mincer, 1974). Mincer relates the concave quadratic function to the behavior that is implied by the optimal distribution of human capital investment over an agent’s life cycle, but more recently Murphy and Welsch (1990) have provided evidence that a quartic relationship performs more satisfactorily than a quadratic one.

Our own application involves a cross-section data set relating to the earnings and ages of a sample of 205 Canadian individuals, all of whom had the same number of years of schooling. These data come from the 1971 Canadian Census Public Use Tapes, and have been used in various parametric and nonparametric studies by Ullah (1985), Singh et al. (1987), Pagan and Ullah (1999, pp. 152–157), and van Akkeren (this volume, Chap. 10). As Pagan and Ullah discuss, simple polynomial models result in a smooth concave relationship, but when a nonparametric kernel estimator is used the fitted relationship exhibits a “dip” around the age of 40–45. Interestingly, the same effect is obtained by van Akkeren (this volume, Chap. 10, Fig. 5) via his data-based information theoretic (DBIT) estimation procedure. A possible explanation for this dip is offered by Pagan and Ullah (1999, p. 154):

[it may be due to] the generation effect, because the cross section data represent the earnings of people at a point in time who essentially belong to different generations. Thus the plot of earnings represents the overlap of the earnings trajectories of different generations. Only if the sociopolitical environment of the economy has remained stable intergenerationally can we assume
these trajectories to be the same. But this is not the case; one obvious counterexample being the Second World War. Therefore, the dip in the nonparametric regression might be attributed to the generation between 1935 and 1945.

Figure 1 illustrates this effect with the Canadian earnings data. There, we show the results of fitting a quadratic relationship by the method of least squares, as well as a nonparametric kernel estimate. The latter uses a normal kernel with the approximately optimal bandwidth described by Silverman (1986, p. 45). Also shown in Fig. 1 is the result of applying our fuzzy regression analysis with three fuzzy clusters and $m = 2$. (The results were not sensitive to the latter choice of value for the exponent parameter in the application of the FCM algorithm.) Some of the details associated with the clustering of the data and the estimation of submodels appear in Table 1, together with a comparison of the quality of the overall “fit” of the fuzzy model as compared with the nonparametric kernel and quadratic least squares models. In Table 1(a) we see that the submodel estimates (based on each of the three fuzzy clusters) are fundamentally different from each other. When these are combined on the basis of the associated membership functions (which are depicted in Fig. 2) we are able to fit a flexible nonlinear relationship. In Table 1(b) we see that the fuzzy model fits the data better than the nonparametric model and

![Figure 1](https://example.com/figure1.png)

**FIGURE 1** Alternative estimates of the earnings-age profile.
TABLE 1 Fuzzy Regression Results for Earning-Age Data \((c = 3; m = 2)\)

(a) Submodel Results

<table>
<thead>
<tr>
<th>Fuzzy cluster</th>
<th>Ranked observations</th>
<th>Cluster centre</th>
<th>Slope ((t\text{-value}))</th>
<th>Intercept ((t\text{-value}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–77 (21–32 years)</td>
<td>25.7421</td>
<td>0.1194 (6.74)</td>
<td>10.1199 (21.68)</td>
</tr>
<tr>
<td>2</td>
<td>78–148 (33–47 years)</td>
<td>39.7241</td>
<td>−0.0180 (−1.31)</td>
<td>14.3870 (26.22)</td>
</tr>
<tr>
<td>3</td>
<td>149–205 (48–65 years)</td>
<td>55.1273</td>
<td>−0.0475 (−2.53)</td>
<td>16.2073 (15.69)</td>
</tr>
</tbody>
</table>

(b) Comparative Model Performance

<table>
<thead>
<tr>
<th></th>
<th>Least squares</th>
<th>Nonparametric</th>
<th>Fuzzy</th>
</tr>
</thead>
<tbody>
<tr>
<td>% RMSE(^a)</td>
<td>4.403</td>
<td>4.459</td>
<td>4.406</td>
</tr>
<tr>
<td>% MAE(^b)</td>
<td>3.017</td>
<td>3.014</td>
<td>2.992</td>
</tr>
</tbody>
</table>

\(^a\)Percentage root mean squared error of fit.
\(^b\)Percentage mean absolute error of fit.

FIGURE 2 Membership functions for fuzzy regression \((c = 3; m = 2)\).
virtually as well as the quadratic least squares model, on the basis of percentage root mean squared error. This quadratic penalty function actually favors least squares, and the fuzzy model emerges as the clear winner when the comparison is based on percentage mean absolute error.

The sensitivity of the fuzzy econometric modeling results to the chosen number of fuzzy clusters is illustrated in Fig. 3, where the results for $c = 3$ and $c = 4$ are compared (with $m = 2$). The latter results appear to “overfit” the data to some degree, and our preference is for the results shown in Fig. 1. The general shape of the fit of the fuzzy regressions reinforces the principal result of the nonparametric estimation (and the DBIT results of van Akkeren, this volume). We see an even more pronounced “dip” than in the case of the nonparametric results, with the minimum occurring at an age of 41 (rather than 44), and the adjacent peaks occurring at ages of 33 and 48 (rather than 38 and 51).

5.2 Aggregate Consumption Function

Our second example involves the estimation of a naive aggregate consumption function for the U.S.A., using monthly seasonally adjusted

FIGURE 3 Sensitivity of fuzzy modeling to number of clusters.
time-series data for personal consumption expenditures and real disposable personal income. The data are published by the Bureau of Economic Analysis at the U.S. Department of Commerce (2000), and are in real 1996 billions of dollars. The sample period is January 1967 to June 2000 inclusive. The model explains consumption expenditure simply as a function of disposable income. No dynamic effects or other explanatory variables are taken into account.

In this case we found that basing the analysis on four fuzzy clusters produced marginally better results than those based on three clusters. Figure 4 compares the “fit” of the fuzzy model with that obtained by simple least squares, and the corresponding comparison is made with a nonparametric kernel fit in Fig. 5. The latter was obtained with the same kernel and window choices as in the previous example, and $m=2$ was used again as the exponent parameter in the application of the FCM algorithm. As can be seen in these two figures, the fuzzy model “tracks” the data much more satisfactorily than do either of its competitors, and this is borne out by the percentage root mean square errors and percentage mean absolute errors that are reported in Table 2. The close similarity between the results based on three and four fuzzy clusters is clear in Fig. 6, and the \% RMSE and \% MAE values when $c=3$ are 1.327 and 1.033,
### Table 2  Fuzzy Regression Results for Consumption–Income Data \( (c = 4; m = 2) \)

#### (a) Submodel Results

<table>
<thead>
<tr>
<th>Fuzzy cluster</th>
<th>Ranked observations</th>
<th>Cluster centre</th>
<th>Slope ((t\text{-value}))</th>
<th>Intercept ((t\text{-value}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–108</td>
<td>2720.247</td>
<td>0.8140 ((92.18))</td>
<td>184.45 ((7.55))</td>
</tr>
<tr>
<td>2</td>
<td>109–217</td>
<td>3681.513</td>
<td>0.8061 ((59.69))</td>
<td>258.91 ((5.19))</td>
</tr>
<tr>
<td>3</td>
<td>218–332</td>
<td>4926.031</td>
<td>0.9740 ((63.76))</td>
<td>/C0 ((-5.26))</td>
</tr>
<tr>
<td>4</td>
<td>333–402</td>
<td>5973.417</td>
<td>1.1705 ((63.52))</td>
<td>/C0 ((-13.05))</td>
</tr>
</tbody>
</table>

#### (b) Comparative Model Performance

<table>
<thead>
<tr>
<th></th>
<th>Least squares</th>
<th>Nonparametric</th>
<th>Fuzzy</th>
</tr>
</thead>
<tbody>
<tr>
<td>% RMSE(^a)</td>
<td>2.306</td>
<td>3.285</td>
<td>1.201</td>
</tr>
<tr>
<td>% MAE(^b)</td>
<td>1.828</td>
<td>2.328</td>
<td>0.942</td>
</tr>
</tbody>
</table>

\(^a\)Percentage root mean squared error of fit.
\(^b\)Percentage mean absolute error of fit.

![U.S. consumption model (monthly 1967M1–2000M6).](image)

**Figure 5**

Giles and Draeseke
The economic interpretation of the various fitted models is also interesting. In the linear least squares model the slope parameter is the (constant) marginal propensity to consume (m.p.c.), and is estimated to be 0.9598 \( (t = 270.39) \) from our sample. In the case of the nonparametric model the estimated derivatives (which represent the changing m.p.c.) range in value from 0.3134 to 1.2427 as the level of disposable income varies. This is shown in Fig. 8, where the income data have been ranked into ascending order. Values of the m.p.c. in excess of unity make no sense in economic terms, and the extreme and unusual pattern of the plot of the nonparametrically estimated m.p.c.’s in Fig. 8 suggests that these results should be treated with extreme skepticism. Also given in that figure are the corresponding results for the preferred Fuzzy (4) model. In this case the plot is much more reasonable than that for the nonparametric model, with the estimated m.p.c.’s ranging from 0.8061 to 1.1705 in value. The few estimates in excess of unity are still troublesome, of course, but apart from these, the derived m.p.c. values are more plausible economically than are those resulting from the nonparametric kernel estimation or from the (highly restrictive) least squares model.
FIGURE 7 Membership functions for fuzzy regression \((c = 4; m = 2)\).

FIGURE 8 Predicted marginal propensities to consume.
5.3 Money Demand Model

Our third empirical application involves a simple demand for money model:

\[
\log(M_t) = \beta_1 + \beta_2 \log(Y_t) + \beta_3 \log(r_t) + \epsilon_t
\]

where \(M\) is the money stock, \(Y\) is the output, and \(r\) is the rate of interest. We have used annual U.S. Department of Commerce (various years) data for the period 1960 to 1983 inclusive, as presented by Griffiths et al. (1993, p. 316). We used three fuzzy clusters to analyze each of the two input variables, and then considered all of the resulting nine combinations of the data partitioning. Of these, five resulted in empty or inadequately sized sets—either there were no sample points consistent with the intersection of the fuzzy sets for \(\log(Y)\) and \(\log(r)\), or there were insufficient to fit a regression. The sample points associated with the remaining four intersections are shown in Table 3, and the TSSS analysis was applied to this four-way partitioning of the data.

As Figs. 9 and 10 indicate, the fit of the fuzzy model clearly dominates that obtained from a nonparametric regression model, and it also dominates an OLS multiple regression analysis—especially at the end-points of the sample. This is confirmed by the % RMSE and % MAE figures that are shown in Table 3(b). Figures 11a and 11b provide details of the membership functions for the income and interest rate input variables, respectively, and Figs. 12a and 12b show the derivatives of the fuzzy model with respect to each of these input variables at each sample point. As the data are all in (natural) logarithms, these derivatives are elasticities, and in these figures they are compared with their counterparts from the nonparametric model. In each case, not only do the fuzzy elasticities evolve more plausibly with an increase in the input variable, but they also have the anticipated signs. In contrast, the elasticities derived from the nonparametric regression model have signs that conflict with the underlying economic theory for some values of income and the interest rate variable. In the case of the OLS regression model based on the full sample period, the (constant) income and interest rate elasticities are 0.7091 \((t = 44.41)\) and \(-0.0533 \((t = -2.49)\), respectively. The sample means of their fuzzy counterparts are quite similar in value, being 0.7368 and \(-0.0606\), respectively, and the unsatisfactory nature of the nonparametric results is underscored when we note that the corresponding sample averages of those estimated elasticities are 0.3237 and 0.0666. (The sign of the latter value, of course, conflicts with the prior theory.) In short, the results of the fuzzy modeling dominate both OLS and nonparametric regression analysis in this example.
5.4 Modeling Kuznets’ “U-Curve”

In a seminal contribution, Kuznets (1955) postulated the existence of an inverted-U relationship between the degree of income inequality and economic growth. He argued that income inequality (perhaps as measured by the Gini coefficient) increases during the early stages of an economy’s growth, reaches a maximum, and then declines as the economy matures. This hypothesis has been subjected to a substantial amount of empirical testing in the development economics literature. The results are somewhat mixed, depending on the type of data used (e.g., time-series, cross-section, or panel), the level of development of the

<table>
<thead>
<tr>
<th>TABLE 3</th>
<th>Fuzzy Regression Results for Money Demand Model ($c_1 = c_2 = 3; m = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) SubModel Results</td>
<td></td>
</tr>
<tr>
<td>Fuzzy cluster</td>
<td>log(Income)</td>
</tr>
<tr>
<td>Observations</td>
<td>Cluster centre</td>
</tr>
<tr>
<td>$Y_1$</td>
<td>1–9</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>10–18</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>19–24</td>
</tr>
<tr>
<td>(b) Cluster Intersection Estimation Results</td>
<td></td>
</tr>
<tr>
<td>Fuzzy intersection</td>
<td>Observations</td>
</tr>
<tr>
<td>$\left(Y_1 \cap R1\right)$</td>
<td>1–6</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>$\left(Y_1 \cap R2\right)$</td>
<td>7–9</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>$\left(Y_2 \cap R2\right)$</td>
<td>10–18</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>$\left(Y_3 \cap R3\right)$</td>
<td>20–24</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) Comparative Model Performances</td>
<td></td>
</tr>
<tr>
<td>Least squares</td>
<td>Nonparametric</td>
</tr>
<tr>
<td>% RMSEb</td>
<td>0.019</td>
</tr>
<tr>
<td>% MAEc</td>
<td>0.015</td>
</tr>
</tbody>
</table>

aNot available, as degrees of freedom are exactly zero.
bPercentage root mean squared error of fit.
cPercentage mean absolute error of fit.
FIGURE 9  U.S. demand for money model.

FIGURE 10  U.S. demand for money model.
FIGURE 11 (a) Membership functions for income variable; (b) Membership functions for interest rate variable.
FIGURE 12  (a) Predicted income elasticities; (b) Predicted interest rate elasticities.
country in question, and the method of estimation. In the case of the U.S.A., the empirical evidence quite clearly rejects Kuznets’ hypothesis in favor of a U-shaped relationship between income inequality and real per capita output. Hsing and Smyth (1994) found support for this result using aggregate U.S. data, whether an allowance was made for ethnic origin or not. Using the same data, but taking into account the nonstationarity of the data, Jacobsen and Giles (1998) also found support for a U-shaped relationship. They also found differences in the output levels at which the minimum of the U-curve occurred for whites as opposed to blacks and others.

Here, we apply our fuzzy modeling to this same data set, focusing simply on the income inequality data for “all” ethnic groups, and abstracting from nonstationarity issues. The relationship that we consider for expository purposes is one in which the only explanatory variable is real per capita GDP. Both Hsing and Smyth (1994) and Jacobsen and Giles (1998) also considered the percentage of married couple families as an additional explanatory variable, and the latter authors also allowed for dynamic effects in the model specification. The benchmark specification that we compare against is a simple but rigid quadratic relationship, estimated by OLS:

$$\text{GA}_t = \beta_1 + \beta_2 \text{GDPPC}_t + \beta_3 \text{GDPPC}_t^2 + \epsilon_t$$

where GA is the (%) Gini coefficient for the entire U.S. population, and GDPPC is per capita real (1987) GDP. The sample period is 1947 to 1991 inclusive. We also estimate a nonparametric kernel regression that explains GA as a function of GDPPC only, and we apply our fuzzy analysis to the model:

$$\text{GA}_t = \beta_1 + \beta_2 \text{GDPPC}_t + \epsilon_t$$

with three fuzzy clusters, and $m = 2$. The results were not sensitive to using four fuzzy clusters, as opposed to three.

The estimation results appear in Table 4, and we again see there that the fuzzy model “fits” the data marginally better than its competitors on the basis of either % RMSE or % MAE. In Fig. 13 we see that the fuzzy model supports the U-shaped relationship found by either OLS or nonparametric regression, though the minimum of the fitted relationship occurs at slightly different output levels in each case. With quadratic OLS estimation this minimum is at a per capita real output level of approximately $13,300, but at approximately $12,700 and $13,500, respectively, in the cases of the fuzzy and nonparametric models. The membership functions associated with the fuzzy clustering of the GDP
TABLE 4  Fuzzy Regression Results for Kuznets’ U-Curve ($c = 3; m = 2$)

(a) SubModel Results

<table>
<thead>
<tr>
<th>Fuzzy cluster</th>
<th>Observations</th>
<th>Cluster centre</th>
<th>Slope ($t$-value)</th>
<th>Intercept ($t$-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–18</td>
<td>10421.87</td>
<td>$-0.5140 \times 10^{-5}$</td>
<td>0.4190 (24.84)</td>
</tr>
<tr>
<td></td>
<td>(1947–1964)</td>
<td></td>
<td>($-3.19$)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19–32</td>
<td>14676.64</td>
<td>$0.2956 \times 10^{-5}$</td>
<td>0.3126 (21.75)</td>
</tr>
<tr>
<td></td>
<td>(1965–1978)</td>
<td></td>
<td>($3.00$)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>33–45</td>
<td>18256.73</td>
<td>$0.9617 \times 10^{-5}$</td>
<td>0.2132 (7.70)</td>
</tr>
</tbody>
</table>

(b) Comparative Model Performances

<table>
<thead>
<tr>
<th></th>
<th>Least squares</th>
<th>Nonparametric</th>
<th>Fuzzy</th>
</tr>
</thead>
<tbody>
<tr>
<td>% RMSE(^{a})</td>
<td>1.574</td>
<td>1.890</td>
<td>1.570</td>
</tr>
<tr>
<td>% MAE(^{b})</td>
<td>1.288</td>
<td>1.595</td>
<td>1.196</td>
</tr>
</tbody>
</table>

\(^{a}\)Percentage root mean squared error of fit.
\(^{b}\)Percentage absolute error of fit.

FIGURE 13  Kuznets’ U-curve model.
data appear in Fig. 14, and the estimated derivatives associated with the fuzzy, OLS, and nonparametric models are plotted in Fig. 15.

5.5 Trend Extraction in Time-Series Data

The fuzzy modeling methodology under discussion in this chapter can, of course, be applied to a wide range of situations. One that is somewhat different from those considered so far is the detection and extraction of the trend component of an economic time series. To illustrate this, we have modeled a relationship of the form:

\[
\log(ER_t) = f(t) + \epsilon_t
\]

where ER is the monthly Canada/U.S. exchange rate (noon spot rate, beginning of month). The logarithmic specification recognizes the likelihood of an underlying multiplicative time-series model. In general, such trend extraction would be based on seasonally adjusted data, but the absence of any discernible seasonal pattern in this particular case makes this unnecessary. Table 5 shows the results of our fuzzy modeling with a seven-cluster specification.
As in the previous examples, we compare the fuzzy modeling with nonparametric kernel estimation. Rather than also consider an OLS model where \( f(t) \) is some polynomial in time, we have used the Hodrick–Prescott (H–P) (1980, 1977) filter as a further benchmark for measuring the trend.

**FIGURE 15** Estimated Gini-GDP derivatives.

**TABLE 5** Fuzzy Regression Results for the Exchange Rate Trend (\( c = 7; m = 2 \))

<table>
<thead>
<tr>
<th>Fuzzy cluster</th>
<th>Observations</th>
<th>Cluster centre</th>
<th>Slope (( t )-value)</th>
<th>Intercept (( t )-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–50</td>
<td>24.07</td>
<td>-0.0010</td>
<td>-0.0273</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-8.38)</td>
<td>(7.70)</td>
</tr>
<tr>
<td>2</td>
<td>51–104</td>
<td>77.37535</td>
<td>0.0032</td>
<td>-0.2054</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(11.51)</td>
<td>(-9.39)</td>
</tr>
<tr>
<td>3</td>
<td>105–158</td>
<td>131.5562</td>
<td>0.0014</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(10.59)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>4</td>
<td>159–212</td>
<td>185.7503</td>
<td>0.0014</td>
<td>0.0171</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(5.16)</td>
<td>(0.33)</td>
</tr>
<tr>
<td>5</td>
<td>213–266</td>
<td>239.6784</td>
<td>-0.0013</td>
<td>0.4860</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-8.00)</td>
<td>(12.16)</td>
</tr>
<tr>
<td>6</td>
<td>267–319</td>
<td>293.1247</td>
<td>0.0020</td>
<td>-0.3126</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(8.37)</td>
<td>(-4.35)</td>
</tr>
<tr>
<td>7</td>
<td>320–368</td>
<td>345.4138</td>
<td>0.0020</td>
<td>-0.3027</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(7.44)</td>
<td>(-3.31)</td>
</tr>
</tbody>
</table>
Our sample covers the period since the (re)floating of the Canadian dollar, and runs from May 1970 to December 2000. We follow Kydland and Prescott (1990, p. 9) and set the smoothing parameter for the H–P filter to \( \lambda = 14,400 \) (i.e., 100 times the square of the data frequency), as we have monthly data. The H–P filter was implemented with use of the TSP package (Hall, 1996) using code written by Cummins (1994).

The membership functions associated with this fuzzy modeling appear in Fig. 16, and Fig. 17 depicts the associated trend analysis. The corresponding nonparametric and H–P filter trend analyses appear in Fig. 18. In this application of H–P filter results are quite robust to the choice of the smoothing parameter. For instance, when \( \lambda = 100 \) (the usual choice of annual data) is used, the results are very similar to those in Fig. 18, but there is undue variation in the extracted trend. On the other hand, in the case of the fuzzy trend analysis a good deal of care has to be taken over the choice of the number of fuzzy clusters if a trend as "flexible" as that produced by the H–P filter is to be obtained. This can be seen in Fig. 19, though obviously this sensitivity is at least in part a function of the sample size and the nature of the trend itself.

In many respects, the trend of these exchange rate data that is identified by means of our fuzzy modeling appears to be somewhat more
FIGURE 17  Exchange rate fuzzy trend analysis.

FIGURE 18  Nonparametric and H-P filter trend analysis.
reasonable than that resulting from the application of the H–P filter. In particular, the latter seems to exhibit too much cyclical variation in this case, at least visually, as can be seen in Fig. 20. A much more comprehensive comparison of these two filtering procedures would be a fruitful research topic, but it is one that is beyond the scope of the present chapter.

However, by way of a simple experiment the following artificial time series has been considered:

\[ y_t = 10 + 5t + 0.2t^2 + 100 \sin(t) + 500 \sin^2(t) + \epsilon_t \]

where \( \epsilon_t \sim N[0, \sigma = 200]; t = 1, 2, 3, \ldots, 100. \)

Figure 21 shows the results of attempting to extract the (known) trend from this series, using a fuzzy model with three membership functions, the H–P filter (with \( \lambda = 100 \)), and nonparametric kernel regression, with “time” as the explanatory variable. As can be seen, the data are extremely variable with very little by way of a discernible trend. The fuzzy trend extractor performs well relative to the other methods, and yields a simple correlation of 0.992 with the true trend component. It performs especially well in the latter part of the sample. The corresponding simple correlations for the H–P filter and the non parametric regression are 0.964 and 0.983, respectively.
FIGURE 20  Fuzzy trend versus H-P trend.

FIGURE 21  Artificial data-comparative trend analysis.
Our final application is based on an example, and data, provided in Studenmund's (1997, pp. 174–175) well known text. This example relates to the demand for chicken, as a function of its own price, disposable income, and the price of a substitute meat. More specifically, the model is of the form:

\[ Q_t = f(P_C^t, P_B^t, Y_t) + \varepsilon_t \]

where:
- \( Q \) = U.S. per capita chicken consumption (pounds)
- \( P_C \) = U.S. price of chicken (cents/pounds)
- \( P_B \) = U.S. price of beef (cents/pound)
- \( Y \) = natural logarithm of U.S. per capita disposable income (dollars)

The sample (Studenmund, 1997, p. 199) comprises annual data for 1951 to 1990 inclusive.

As in the earlier regression examples, we considered OLS, nonparametric kernel estimation, and fuzzy modeling. The last was applied to all three (nonconstant) explanatory variables. In each case we

---

**FIGURE 22** Chicken consumption model.
used two fuzzy clusters, resulting in eight potential subsamples, three of which had positive degrees of freedom. (Three fuzzy clusters implied 27 potential subsamples, and this resulted in only two with positive degrees of freedom. The resulting fuzzy prediction path exhibited erratic movement at several time points.)

The basic results are shown in Fig. 22 and Table 6. The membership functions were of the same general form as in the previous examples, and are not shown here to reduce space. As can be seen in Fig. 22, the fuzzy model performs very creditably, and this is reflected in the % RMSE and % MAE values shown in Table 6. Indeed, if it were not for the relatively poor predictive performance of the fuzzy model during the period 1974–1981, its overall performance would be exceptionally good. Our view is

**Table 6** Fuzzy Regression Results for Chicken Consumption Model

(c = c2 = c3 = 2; m = 2)

(a) Submodel Results

<table>
<thead>
<tr>
<th>Price of chicken</th>
<th>Price of beef</th>
<th>log(Income)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuzzy cluster</td>
<td>Cluster centre</td>
<td>Fuzzy cluster centre</td>
</tr>
<tr>
<td>$P^C_1$</td>
<td>10.420</td>
<td>$P^B_1$</td>
</tr>
<tr>
<td>$P^C_2$</td>
<td>18.166</td>
<td>$P^B_2$</td>
</tr>
</tbody>
</table>

(b) Cluster Intersection Results

<table>
<thead>
<tr>
<th>Fuzzy intersection</th>
<th>Observations</th>
<th>$\beta_1$ (t-value)</th>
<th>$\beta_2$ (t-value)</th>
<th>$\beta_3$ (t-value)</th>
<th>$\beta_4$ (t-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>($P^C_1 \cap P^B_1 \cap Y_1$)</td>
<td>7–22</td>
<td>30.4720</td>
<td>-1.0523</td>
<td>0.2901</td>
<td>0.0026</td>
</tr>
<tr>
<td></td>
<td>24–27</td>
<td>(9.90)</td>
<td>(-5.34)</td>
<td>(1.84)</td>
<td>(3.70)</td>
</tr>
<tr>
<td>($P^C_1 \cap P^B_2 \cap Y_2$)</td>
<td>29–33</td>
<td>9.7358</td>
<td>-0.0243</td>
<td>0.3360</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td>36–38, 40</td>
<td>(1.42)</td>
<td>(-0.07)</td>
<td>(4.29)</td>
<td>(13.36)</td>
</tr>
<tr>
<td>($P^C_2 \cap P^B_2 \cap Y_1$)</td>
<td>1–6</td>
<td>15.1140</td>
<td>-0.2595</td>
<td>0.1433</td>
<td>0.0061</td>
</tr>
<tr>
<td></td>
<td>15–18</td>
<td>(0.70)</td>
<td>(-0.91)</td>
<td>(0.82)</td>
<td>(0.58)</td>
</tr>
</tbody>
</table>

(c) Comparative Model Performances

<table>
<thead>
<tr>
<th></th>
<th>Least squares</th>
<th>Nonparametric</th>
<th>Fuzzy</th>
</tr>
</thead>
<tbody>
<tr>
<td>% RMSE(^b)</td>
<td>5.130</td>
<td>7.692</td>
<td>6.879</td>
</tr>
<tr>
<td>% MAE(^{c})</td>
<td>4.407</td>
<td>6.001</td>
<td>4.559</td>
</tr>
</tbody>
</table>

\(^a\) $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ are the coefficients of the intercepts, $P^C$, $P^B$, and $Y$ respectively.

\(^b\) Percentage root mean squared error of fit.

\(^c\) Percentage mean absolute error of fit.
that the relatively small sample size disadvantages the fuzzy model in this example, and this may also be an instance where some experimentation is needed with the value for the exponent parameter, \( m \), in the objective functional for the FCM algorithm.

Figure 23 displays the predicted own price, cross price, and income elasticities of demand (by year in the sample) implied by the OLS, nonparametric, and fuzzy models. Those relating to the nonparametric estimation are quite unsatisfactory. In particular, in each case there are values that have a sign opposite to that implied by the underlying economic theory. There are no such aberrations in the case of the fuzzy and OLS results. The latter generally appear somewhat more plausible as they exhibit fewer marked fluctuations than do the fuzzy elasticities. This weakness of the fuzzy elasticities is related to a similar feature in the predicted time paths, noted above in relation to Fig. 22. It is also interesting that the fuzzy income elasticities always exceed those from OLS estimation, and this is almost always the case with the cross-price elasticities.

**FIGURE 23** (a) Predicted own-price elasticities; (b) Predicted cross-price elasticities; (c) Predicted income elasticities (multiplied by 10,000).
FIGURE 23  (Continued)
6 CONCLUDING REMARKS

In this chapter we have discussed the possibility of using various analytical tools from fuzzy set theory and fuzzy logic to model nonlinear econometric relationships in a flexible, and essentially semiparametric way. More specifically, we have explored the use of the fuzzy c-means clustering algorithm, in conjunction with the Takagi and Sugeno (1985) approach to fuzzy systems, in this particular context. The general modeling strategy that we have considered involves identifying interesting and important fuzzy sets or fuzzy clusters of the multidimensional data, in a totally nonparametric way; fitting separate parametric regression models to each fuzzy cluster (or sub-sample) of the data; and then combining the estimates of these separate models in a very flexible way, based on the “degree of membership” of each sample point to each fuzzy cluster. It is this last step in the analysis that facilitates the overall procedure’s ability to capture intrinsic nonlinearities, because the “weights” that are effectively assigned to each submodel vary continuously from data point to data point.

This general fuzzy modeling procedures has a wide range of applications, and we have provided several empirical applications to illustrate this. Not only can it be used to model and smooth data in ways that compete directly with established techniques such as nonparametric kernel regression, or splines, but it can also be used for trend extraction in competition with methods such as the Hodrick and Prescott (1980, 1997) filter. In all of the cases that we have examined, the fuzzy modeling approach performs extremely well. Moreover, it has significant practical advantages over both kernel regression and spline analysis. Spline analysis requires that the location of the “knots” be known in advance, whereas the first stage of our fuzzy modeling procedure determines the relevant data groupings empirically. Nonparametric kernel regression suffers from the so-called “curse of dimensionality” the severely limits its application to relatively simple models. In contrast, the fuzzy modeling structure that we have outlined is readily applicable to quite complex models, without undue computational burden. Indeed, it should also be noted that although our illustrative applications have all used ordinary least squares regression at the second stage of the analysis to fit models to each fuzzy cluster, in fact any relevant estimation technique could be used at that stage. For instance (and depending on the context), logit or probit models could be fitted, or nonlinear or instrumental variables regression could be used, without disrupting the general style of the fuzzy modeling strategy that we have described.

We recognize that the bulk of the discussion in this chapter addresses issues relating to curve fitting and data smoothing, and relatively little
attention has been paid to strictly inferential issues. It was argued
heuristically at the end of Section 4 that the fuzzy predictor emerging from
our modeling approach will be a (weakly) consistent estimator of the
conditional mean of the dependent variable. While this is minimally
helpful, it will hold only under suitable conditions, and our work in
progress explores this issue (and other aspects of the sampling properties
of the fuzzy predictor) more thoroughly.

In this context it is also worth noting that the various "fuzzy
predictions" that we have presented in our examples are simply "point
predictions," and the corresponding confidence intervals are also extremely
important. While it is not immediately clear how straightforward it would
be to construct exact finite-sample prediction intervals within this
framework, they could certainly be approximated by means of bootstrap
simulation. Asymptotic prediction intervals can be constructed relatively
easily, as follows. It will be recalled that once the sample has been
partitioned into fuzzy clusters (typically with a different number of data
points in each cluster), our modeling procedure involves estimating a
regression over each cluster, and then combining the results using weights
based on the membership values. The precise details are given in Step 3 in
Section 4. These weights vary continuously, but are exogenous if the
regressors also satisfy this property. Rather than estimate each cluster
regression separately, by least squares, another option is to estimate the

FIGURE 24 Confidence intervals for earning–age profile predictions.
group of cluster regressions as a “seemingly unrelated regression” (SUR) model. The only complication is that the SUR model is “unbalanced,” in the sense that the samples associated with each equation are different from one another. However, unbalanced SUR models can be estimated quite readily (e.g., see Srivastava and Giles, 1987, pp. 339–346, for details). This then provides a complete estimated asymptotic covariance matrix for all of the estimated parameters in all the cluster regressions. This information (together with the membership values) can then be used to construct observation-by-observation asymptotic standard errors for the prediction of the conditional mean of the dependent variable.

By way of illustration we have undertaken this analysis in the case of the earnings–age profile data analyzed in Section 5.1, and the results appear in Fig. 24. SHAZAM code was written to produce these results, which are compared with their counterparts from nonparametric kernel regression. As can be seen, the fuzzy modeling procedure produces intervals that are very comparable to those from nonparametric regression, in terms of their shape and width. Given that our illustrative applications are based on relatively small samples, we have not provided corresponding asymptotic confidence intervals for our other “fuzzy predictions.”

Clearly, much remains to be done in order to validate the real worth of the type of “fuzzy econometric modeling” that has been described and illustrated in this chapter. Nonetheless, we view the results presented here as being rather promising. Certainly, further exploration of this approach to modeling and smoothing nonlinear economic data seems to be justified.

ACKNOWLEDGMENTS

We are grateful to George Judge, Joris Pinkse, Ken White, and participants in the University of Victoria Econometrics Colloquium for their helpful comments.

REFERENCES


1 INTRODUCTION
Since the path-breaking work of Karl Pearson the 20th century saw significant advance in parametric statistical and econometric hypothesis-testing procedures [see Bera (2000) for an excellent survey]. A problem with the parametric testing procedures is that the tests may not be consistent under the mis-specified alternative hypotheses. In the last two decades a rich literature has developed on constructing consistent model specification tests using nonparametric estimation techniques. Bierens (1982) was the first to provide a consistent conditional moment test for model mis-specification. Ullah (1985) first suggested the construction of a model specification test using the nonparametric estimation technique. A nonparametric specification test for time-series data was first proposed by Robinson (1989).


While most of the early developments in nonparametric hypothesis testing appeared for the independent data, except, e.g., Robinson (1989), in recent years the problem of hypothesis testing with the dependent time-series data has been addressed by many authors. For example, Berg and Li (1998), Fan and Li (1997), Hjellvik and Tjøstheim (1995, 1998), Hjellvik et al. (1998), Kreiss (1998), Lee (2001), and Lee and Ullah (2001), among others, have considered tests for functional form and omitted variables. In particular, for time series, testing for omitted variables is often to identify the number of lags. Nonparametric lag selection in nonlinear time-series models are studied by Auestad and Tjøstheim (1990), Cheng and Tong (1992), Fan and Li (1999a), Granger and Lin (1994), Granger et al. (2000), Hong and White (2001), Tjøstheim and Auestad (1994), Tschernig and Yang (2000), and Yao and Tong (1994). Chen and Fan (1999) provided consistent tests for time-series models, but the asymptotic distributions of their tests were nonstandard. In a major development, Fan and Li (1999b) developed the central limit theorems for the degenerate U-statistics for the weakly dependent data. This has provided a significant breakthrough and important contribution in Li (1999), who shows the asymptotic normality of Li and Wang (1998) type tests for a wide range of hypothesis-testing problems, with dependent data, e.g., parametric functional forms in regression, single index models, semiparametric regressions, variable selection, and mean–variance ratio hypothesis in finance.

We note here that the test statistics for most of the testing problems described above are based on the following alternative procedures: (1) Ullah (1985) type F or likelihood ratio procedure comparing the residual sum of squares under the null and alternative hypotheses (see also
Azzalini et al., 1989; Cai et al., 2000; Fan et al., 2001); (2) the procedure comparing the sum of squares of the differences in the fitted values of the models under the null and alternative specifications, e.g., Härdele and Mammen (1993), Ullah and Vinod (1993), and Aıt-Sahalia et al. (1994); and (3) the conditional moment procedure looking into the covariance between the residual under the null and the model specified under the alternative (e.g., Fan and Li, 1996; Zheng, 1996; Li and Wang, 1998). These three alternative procedures are equivalent in the sense that they conform to the same population value of the null hypothesis of no difference between the null and alternative specifications. However, sample statistics based on them are different and may give different results. The purpose here is not to introduce a new procedure of nonparametric testing, instead the modest aim is to explore the bootstrap simulation comparison of these three procedures with respect to size and power properties in small as well as large samples. In the earlier simulation studies usually the case of testing a parametric specification is considered by only one of the three procedures. This is perhaps the first study that considers all the three procedures and looks into not only the testing of a parametric specification but also the testing of varying parameters and omitted variables. We also consider both the independent cross-section and dependent time-series data models. Both the naive bootstrap and wild bootstrap procedures are used for out analysis.

The plan of the chapter is as follows. In Section 2, we present the nonparametric kernel regression estimator. Section 3 presents the three procedures of nonparametric hypothesis testing. Then, in Section 4 we give our simulation results. Finally, Section 5 gives conclusions.

2 NONPARAMETRIC REGRESSION

Let us consider the regression model:

\[ y_t = m(x_t) + u_t = E(y_t|x_t) + u_t \]

where \( t = 1, \ldots, n \), \( y_t \) is a scalar dependent variable, \( x_t = (x_{t1}, \ldots, x_{tk}) \) is an \( 1 \times k \) vector, \( m(x_t) = E(y_t|x_t) \) is the true but unknown regression function, and \( u_t \) is the error term such that \( E(u_t|x_t) = 0 \). The model in Eq. (1) includes the autoregressive model as a special case in which \( x_t \) consists of lagged values of \( y_t \). For the time-series case we assume that \( \{y_t, x_t\} \) is a strictly stationary discrete-time stochastic process.

A parametric approach to estimating \( m(x_t) \) in Eq. (1) may begin by fitting a linear parametric regression model through the data as

\[ y_t = \alpha + \beta x_t + u_t \]

\[ = X_t \delta + u_t \]
or more generally a nonlinear parametric model \( y_t = m(x_t, \delta) + u_t \), where \( X_t = (1, x_t) \) and \( \delta = (\alpha \beta)' \). One can obtain a least squares (LS) estimator of \( \hat{m}(x_t) \) by \( m(x_t, \hat{\delta}) \) where \( \hat{\delta} \) is the LS estimator of \( \delta \) obtained by minimizing the global LS objective function:

\[
\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y - m(x_i, \delta))^2
\]  

However, this global parametric LS estimator, based on global modeling, is inconsistent and biased at least in the regions of data where the a priori specified regression is not correctly specified.

An alternative improved approach is to use the nonparametric kernel regression estimation of the unknown \( m(x_t) \). Essentially, the idea behind the kernel regression is to model the regression function \( m(x_t) \) locally. For example, to obtain the regression function at a given point \( x \), we apply the standard linear regression technique to the data in the interval of length \( h \) around \( x \). That is, for the data in the interval of length \( h \), we consider the linear model:

\[
y_t = \alpha(x) + x_t \beta(x) + u_t = X_t \beta(x) + u_t
\]  

and then estimate \( \beta(x) \) by minimizing the local LS or weighted LS errors:

\[
\sum_{i=1}^{n} u_i^2 K_{tx} = \sum_{i=1}^{n} (y_i - X_t \beta(x))^2 K_{tx}
\]  

with respect to \( \beta(x) \), where \( K_{tx} = K((x_t - x)/h) \) is called a kernel (weight) function and \( h \to 0 \) as \( n \to \infty \) is usually called the window width (smoothing parameter). Generally, the kernel function can be any probability density function having a finite second moment. The estimator so obtained is

\[
\hat{\beta}(x) = (X'K(x)X)^{-1}X'K(x)y
\]  

where \( K(x) \) is the \( n \times n \) diagonal matrix with diagonal elements \( K_{tx}(t=1, \ldots, n) \), \( X \) is an \( n \times (k+1) \) matrix with the \( t \)th row \( X_t \), and \( y \) is an \( n \times 1 \) vector. The estimator of \( m(x_t) \) is then given by \( m(x_t) = X_t \hat{\beta}(x) \).

The approach in Eq. (4) is called local linear regression and the estimator in Eq. (6) is known as the local linear LS (LLLS) estimator. For details on the kernel regression estimators and the choices of \( K(\cdot) \) and \( h \), see Wand and Jones (1995), Fan and Gijbels (1996), and Pagan and Ullah (1999).

It is interesting to note the special cases and generalization of Eqs. (4) and (6). When \( h = \infty \), the local linear regression modeling in Eq. (4) becomes global modeling in Eq. (2) and the LLLS estimator \( \hat{\beta}(x) \) in Eq. (6)
becomes the global LS estimator of $\hat{\delta}$. This is because when $h = \infty$, $K_{tx} = K(0)$ and the minimization of $\Sigma(y_t - X_t \delta(x))^2$ becomes the minimization of $K(0) \Sigma(y_t - X_t \delta(x))^2 = \text{minimization of } \Sigma(y_t - X_t \delta)^2$. Also note that when $X_t = 1$, the LLLS estimator $\hat{\delta}(x)$ reduces to

$$\hat{\delta}(x) = \hat{\delta}(x) = (i' K(x)i)^{-1} i' K(x)y$$

which is the Nadaraya (1964) and Watson (1964) kernel regression estimator, where $i$ is an $n \times 1$ vector of unit elements. The LLLS can be extended to the $p$th order local polynomial LS estimator where, for $k = 1$, $X_t = [1, x_t, \ldots, x_t^p]$ and $\delta$ is a $(p + 1) \times 1$ vector (see Fan and Gijbels, 1996).

One advantage of the local estimators is that they can be viewed as the varying coefficient (functional coefficient) estimators. This is because $\hat{\delta}(x)$ may have varying values at different data points $x_t$. In this sense the local linear model $y_t = X_t \delta(x) + u_t$ is a varying coefficient model $y_t = m(x_t) + u_t$ where $m(x_t) = X_t \delta(x) \approx X_t \delta(x)$ where $\delta(x)$ is the first term of the Taylor's approximation around $x$. This is in contrast to the global estimator $\hat{\delta}$, which is the estimator of $\delta$ in the constant coefficient model.

The above idea of a varying coefficients model can be extended to the situations where the coefficients are varying with respect to $z_t$, which may be a subset of $x_t$ or something else, i.e.,

$$E(y_t|x_t) = m(x_t) = X_t \delta(z_t)$$

Examples of these include the functional coefficient autoregressive model (Chen and Tsay, 1993; Cai et al., 2000), smooth coefficient model (Li et al., 1997), random coefficient model (Raj and Ullah, 1981), smooth transition autoregressive model (Granger and Teräsvirta, 1993), exponential autoregressive model (Haggan and Ozaki, 1981), and threshold autoregressive model (Tong, 1990). Also see Section 4. To estimate $\delta(z_t)$ we can again do a local approximation $\delta(z_t) \approx \delta(z)$ and then minimize $\Sigma(y_t - X_t \delta(x))^2 K_{tx}$ with respect to $\delta(z)$, where $K_{tx} = K((z_t - z)/h)$. This gives the varying coefficient estimator:

$$\hat{\delta}(z) = (X K(z)X)^{-1} X K(z)y$$

where $K(z)$ is a diagonal matrix of $K_{tz_t}$, $t = 1, \ldots, n$. When $z_t = x_t$, this reduces to the LLLS estimator $\hat{\delta}(x)$ in Eq. (6).

Cai et al. (2000) consider a local linear approximation $\delta(z_t) \approx \delta(z) + D(z)(z_t - z)'$ where $D(z) = (\partial \delta(z)/\partial z_t)'$ evaluated at $z_t = z$. The LL
varying coefficient (LLVC) estimator of Cai et al. (2000) is then obtained by minimizing

\[ \sum_{i=1}^{n} \left[ y_{i} - X_{i} \delta(z_{i}) \right]^{2} K_{z_{i}} \]

\[ = \sum_{i=1}^{n} \left[ y_{i} - X_{i} \delta(z) - \left[ (z_{i} - z) \otimes X_{i} \right] \text{vec } D(z) \right]^{2} K_{z_{i}} = \sum_{i=1}^{n} \left[ y_{i} - X_{i} \delta(z) \right]^{2} K_{z_{i}} \]

with respect to \( \delta(z) = [\delta(z)^{T} \ (\text{vec } D(z))^{T}] \) where \( X_{i} = [X_{i} (z_{i} - z) \otimes X_{i}] \). This gives

\[ \delta(z) = (X^{T} K(z) X)^{-1} X^{T} K(z) y \]  

(9)

and \( \delta(z) = (0 0)^{T} \). Hence,

\[ \tilde{m}(x) = (1 0)^{T} \delta(z) = (1 x) \tilde{\delta}(z) \]  

(10)

For the asymptotic properties of these varying coefficient estimators, see Cai et al. (2000).

3 NONPARAMETRIC BOOTSTRAP TESTS

We consider here two types of null hypotheses on \( m(\cdot) \):

\[ H_{0} : m(x_{i}) = m(x_{i}, \delta) \]  

(11)

\[ H_{0} : m(x_{i}) = m(x_{i1}) \]  

(12)

where \( x_{i} = (x_{i1}, x_{i2}) \); \( x_{i1} \) is \( 1 \times k_{1} \), \( x_{i2} \) is \( 1 \times k_{2} \), and \( k = k_{1} + k_{2} \). The alternative hypothesis in each case is the unspecified nonparametric regression:

\[ H_{1} : m(x_{i}) = E(y_{i} | x_{i}) \]  

(13)

The null hypothesis in Eq. (11) can be used as the null hypothesis for testing both for the functional form as well as the varying coefficient models. As a simple example of Eq. (11), we may consider testing for the functional form to be a linear regression, namely, \( H_{0} : m(x_{i}, \delta) = X_{i} \delta \) against \( H_{1} \) in Eq. (13). However, if \( H_{1} \) in Eq. (13) is specified to be local linear models \( m(x_{i}) = X_{i} \delta(x_{i}) \) or \( m(x_{i}) = X_{i} \delta(z_{i}) \) then the testing for the null \( H_{0} : m(x_{i}, \delta) = X_{i} \delta \) against \( H_{1} : m(x_{i}) = X_{i} \delta(z_{i}) \) is testing for the constant regression model against the varying coefficient regression (see Section 4 for more examples).

The null hypothesis in Eq. (12) is for testing the significance of the omitted variables \( x_{i2} \), i.e., to test for selection of variables or lags. In this case both the null and alternative models are nonparametric. Though not
considered here the test statistics considered below can also be used for testing situations where the null hypothesis is the partially linear model, \( m(x_i) = x_i \beta + m(x_{i2}) \), single index models \( m(x_i) = m(x_i, \delta) \), among others.

We will consider three main approaches for the above testing problems.

The first test procedure we consider is, as suggested in Ullah (1985), to compare the residual sum of squares RSS\(_0\) under the null with the nonparametric residual sum of squares under the alternative, RSS\(_1\). The test statistic is

\[
T = \frac{(RSS_0 - RSS_1)}{RSS_1}
\]

where for the null, Eq. (11), RSS\(_0\) = \( \sum \hat{u}_t^2 \), \( \hat{u}_t = y_t - m(x_i, \delta) \), and for the null, Eq. (12), RSS\(_0\) = \( \sum \tilde{u}_t^2 \), \( \tilde{u}_t = y_t - \hat{m}(x_{i1}) \), and \( \hat{m}(x_{i1}) \) is given by Eq. (10) with \( x_i = x_{i1} \). Further, RSS\(_1\) = \( \sum \tilde{u}_t^2 \), \( \tilde{u}_t = y_t - \tilde{m}(x_i) \). We reject the null hypothesis when \( T \) is large.

Fan and Li (1992) show the asymptotic normality of \( nh^{1/2}T \) [see also Cai et al. (2000) and Fan et al. (2001, Theorem 5)].

Fan et al. (2001) further show that a suitably normalized \( T \) will have its asymptotic null distribution that is independent of nuisance parameters. They call this property the Wilks (1938) phenomenon. An important consequence of this result is that one does not have to derive theoretically the normalizing factors in order to be able to use the test. As long as the Wilks phenomenon holds, one can simply simulate the null distribution of the test statistic \( T \). This is in stark contrast with some other tests whose asymptotic null distributions depend on nuisance parameters. Based on these Wilks results of Fan et al. (2001), Cai et al. (2000) suggest using the bootstrap method, which allows the implementation of Eq. (14). It involves the following steps to evaluate \( p \) values of \( T \) to test the null hypotheses in Eqs. (11) and (12):

1. Generate the bootstrap residual \( \{\tilde{u}_t^*\} \) from the centered residuals from the nonparametric (NP) alternative model \( \{\tilde{u}_t - \hat{u}_t\} \) where \( \hat{u} = \frac{1}{n} \sum \hat{u}_i \) and \( \hat{u}_t = y_t - \hat{m}(x_i) \).
   a. For naive bootstrap, \( \{\tilde{u}_t^*\} \) is obtained from randomly resampling \( \{\tilde{u}_t - \hat{u}_t\} \) with replacement.
   b. For wild bootstrap, \( \tilde{u}_t^* = a(\tilde{u}_t - \hat{u}_t) \) with probability \( r = (\sqrt{5} + 1)/2\sqrt{5} \) and \( \tilde{u}_t^* = b(\tilde{u}_t - \hat{u}_t) \) with probability \( 1 - r \times (t = 1, \ldots, n) \), where \( a = -(\sqrt{5} - 1)/2 \) and \( b = (\sqrt{5} + 1)/2 \). See Li and Wang (1998, pp. 150–151).
2. Generate the bootstrap sample \( \{\tilde{y}_t^*\}_{t=1}^n \) from the null model, from \( \tilde{y}_t^* = m(x_i, \delta) + \tilde{u}_t^* \) \( (t = 1, \ldots, n) \) for the null in Eq. (11) to test for
parametric functional form, and from $y_t^* = m(x_{t1}) + \tilde{u}_t^*$ for the null in Eq. (12) to test for omitted variables $x_{t2}$.

3. Using the bootstrap sample $\{y_t^*, x_t, z_t\}_{t=1}^n$, calculate the bootstrap test statistic $T^*$ using, for the sake of simplicity, the same $h$ used in estimation with the original sample as done in Cai et al.

4. Repeat the above steps $B$ times and use the empirical distribution of $T^*$ as the conditional null distribution of $T$ given $\{y_t^*, x_t, z_t\}_{t=1}^n$. We use $B = 500$. The bootstrap $p$ value of the test $T$ is simply the relative frequency of the event $\{T^* \geq T\}$ in the bootstrap resamples.

We use both naive bootstrap (Efron, 1979) and wild bootstrap (Wu, 1986; Liu, 1988). The wild bootstrap method preserves the conditional heteroskedasticity in the original residuals. For wild bootstrap, see also Härdle (1990, p. 247), Shao and Tu (1995, p. 292), or Li and Wang (1998, p. 150).

Two more versions of the $T$ test in (14) can be considered:

\[ S = \frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_0} \]  

\[ R = \frac{1}{n}(\text{RSS}_0 - \text{RSS}_1) \]  

where $S$ is the same as $T$ with $\text{RSS}_1$ in the denominator replaced by $\text{RSS}_0$ in the spirit of Rao’s score test, and $R$ is essentially the numerator of $T$.

In our Monte Carlo study in Section 4 the statistics $T, S,$ and $R$ are compared and calculated on the basis of weighted (trimmed) RSS to control the tail behavior of the nonparametric estimator. For example, weighted $\text{RSS}_1 = \Sigma \tilde{w}(z_t, a)$, and for the null in Eq. (11) weighted $\text{RSS}_0 = \Sigma \tilde{w}(z_t, a)$, where $\tilde{w}(z_t, a) = I(|z_t/\hat{\sigma}_z| < a)$, $\hat{\sigma}_z$ is the sample standard deviation, $a = \infty, 2, 1.5$, and $I(\cdot)$ is the indicator function. The test statistics with weighted RSSs will be denoted as $R_a, S_a,$ and $T_a$. Note that $\tilde{w}(z_t, \infty) = 1$ and thus $R_\infty, S_\infty, T_\infty$ are $R, S,$ and $T$ in Eqs. (14)–(16) without weights. When the weight $\tilde{w}(\cdot)$ is a known function so that it is not estimated, the Wilks phenomenon continues to hold as shown by Fan et al. (2001, Remark 4.2 and Theorem 9) and thus the Cai et al. (2000) bootstrap procedure can be applied to the statistics $R_a, S_a,$ $T_a$ with weighted RSSs.

The second test procedure we consider is to compare the fitted values from the null and alternative models as suggested in Härdle and Mammen.
(1993), Ullah and Vinod (1993), and Aı ¨t-Sahalia et al. (1994). For the null in Eq. (11), this is given by

\[ Q_a = \frac{1}{n} \sum_{t=1}^{n} (m(x_t, \hat{\delta}) - \hat{m}(x_t))^2 w(z_t, a) \]  

(17)

For the omitted variable testing in Eq. (12), \( m(x_t, \hat{\delta}) \) in Eq. (17) is replaced by the NP estimator \( \hat{m}(x_t) \) with \( x_{t2} \) omitted. The bootstrap procedure described above for \( R_a, S_a, T_a \) may also be applied to \( Q_a \).

The third test procedure we consider is the conditional moment test for \( E(ut|xt) = 0 \), which is identical to testing \( E(ut) f(x_t) / C_1 = 0 \) where \( f(x_t) \) is the density of \( x_t \). A sample estimator of the left-hand side of Eq. (18) is

\[ L'_0 = \frac{1}{n} \sum_{t=1}^{n} \hat{u}_t E(\hat{u}_t | x_t) \hat{f}(x_t) \]

\[ = \frac{1}{n(n-1)h^2} \sum_{t=1}^{n} \sum_{t' \neq t} \hat{u}_t \hat{u}_{t'} K_{t't} \]  

(19)

where \( \hat{u}_t = y_t - m(x_t, \hat{\delta}) = y_t - X_t \hat{\delta} \) to test for the null hypothesis in Eq. (11) or \( \hat{u}_t = y_t - \hat{m}(x_t) \) to test for the null hypothesis in Eq. (12), \( E(\hat{u}_t | x_t) = \sum_{t' \neq t} \hat{u}_t K_{t't} / \sum_{t' \neq t} K_{t't} \) from Eq. (7), and \( \hat{f}(x_t) = [(n-1)h^2]^{-1} \sum_{t' \neq t} K_{t't} \) is the kernel density estimator, \( K_{t't} = K((x_t - x_{t'})/h) \).

Note that we estimate the auxiliary regression function \( E(\hat{u}_t | x_t) \) from the local constant LS estimator, Eq. (7), of Nadaraya (1964) and Watson (1964), not from the LLVC estimator of Cai et al. (2000) in Eq. (9) just to maintain the original formula of Li and Wang (1998) and Zheng (1996).

The asymptotic test statistic is then given by

\[ L = nh^{k/2} \frac{L'}{\sqrt{\hat{\sigma}}} \xrightarrow{d} N(0,1) \]  

(20)

where \( \hat{\sigma} = 2(n(n-1)h^2)^{-1} \sum_{t' \neq t} \hat{u}_t^2 \hat{u}_{t'}^2, K_{t't}^2 \) is a consistent estimator of the asymptotic variance of \( nh^{k/2} L' \); see Fan and Li (1996), Zheng (1996), Li and Wang (1998), Fan and Ullah (1999), and Rahman and Ullah (1999) for details. Also, see Pagan and Ullah (1999, Chap. 3) and Ullah (2001) for the relationship between \( R, Q, \) and \( L \) test statistics. Based on the asymptotic results of Fan and Li (1996, 1997, 1999b) and Li (1999) for dependent data, Berg and Li (1998) establish the asymptotic validity of using the wild bootstrap method for \( L \) for time series. The bootstrap \( p \)
values for $L$ to test for the adequacy of the linear parametric model, $m(x, \delta) = X\delta$ in Eq. (11), can be computed as follows:

1. Generate the bootstrap residuals $\{\hat{u}_t^*\}$ from the residual from the null model $\hat{u}_t = y_t - \hat{X}\hat{\delta}$:
   a. For naive bootstrap, $\{\hat{u}_t^*\}$ is obtained from randomly resampling $\{\hat{u}_t\}$ with replacement.
   b. For wild bootstrap, $\hat{u}_t^* = a\hat{u}_t$ with probability $r$ and $\hat{u}_t^* = b\hat{u}_t$ with probability $1-r$ as discussed above.
2. Generate the bootstrap sample $\{y_t^*_n\}_{t=1}^n$ from the null model $y_t^* = m(x_t, \hat{\delta}) + \hat{u}_t^* = X_t\hat{\delta} + \hat{u}_t^*(t = 1, \ldots, n)$ for the null in Eq. (11) to test for neglected nonlinearity.
3. Using the bootstrap sample $\{y_t^*_n\}_{t=1}^n$, calculate the bootstrap test statistic $L^*$.
4. Repeat the above steps $B$ times and use the empirical distribution of $L^*$ as the null distribution of $L$. We use $B = 500$. The bootstrap $p$ value of the test $L$ is the relative frequency of the event $\{L^* \geq L\}$ in the bootstrap resamples.

For testing the omitted variables in the null, Eq. (12), we replace $X_t\hat{\delta}$ in the steps 1 and 2 above by the NP regression estimator $\tilde{m}(x_{t1})$ in Eq. (10) with $x_t = x_{t1}$ and with $x_{t2}$ omitted, and use centered NP residuals. That is:

1. Generate the bootstrap residuals $\{\tilde{u}_t^*\}$ from the centered residuals from the NP alternative model $(\hat{u}_t - \hat{u})$ where $\hat{u} = n^{-1}\sum \hat{u}_t$ and $\tilde{u}_t = y_t - \tilde{m}(x_{t1})$.
2. Generate the bootstrap sample $\{y_t^*_n\}_{t=1}^n$ from the null model $y_t^* = \tilde{m}(x_{t1}) + \tilde{u}_t^*$.

For parametric models, Davidson and MacKinnon (1999) show that the size distortion of a bootstrap test is at least of the order $n^{-1/2}$ smaller than that of the corresponding asymptotic test. For nonparametric models, $h$ also enters in the order of refinement. Li and Wang (1998) show that if the distribution of $L_j^*$ ($j = A$ for asymptotic, $B$ for naive bootstrap, and $W$ for wild bootstrap) admit an Edgeworth expansion then the bootstrap distribution approximates the null distribution of $L$ with an error of order $n^{-1/2}h^{3/2}$, improving over the normal approximation. Since $L$ is asymptotically normal under the null, the bootstrap tests $L_B^*$ and $L_W^*$ are more accurate than the asymptotic test $L_A^*$, as confirmed in the simulation of Section 4. See Hall (1992) for further discussion on Edgeworth expansions and the extent of the refinements in various contexts.

The above three different testing approaches are related. Under the null in Eq. (11) $H_0: m(x_t) = m(x, \delta)$, the RSS-based test statistics $R, S, T$ will
be expected to be zero as $E(y_t - m(x_t))^2 = E(y_t - m(x_t))^2$. Also, by construction $E(u_t|x_t) = 0$ under the null, which implies $E[u_tE(u_t|x_t)f(x_t)] = 0$ for the $L$ test. From this, we obtain the relationship used for the $Q$ test because $E[u_tE(u_t|x_t)f(x_t)] = E[E(y_t|x_t) - m(x, \delta)]^2f(x_t) = 0$.

4 MONTE CARLO

In this section we examine the finite sample properties of the test statistics $T, Q, L$ especially with the empirical null distributions being generated by the bootstrap method. Asymptotic critical values are also used for the $L$ test. We consider four cases as indicated in “blocks” below. All of the error terms $\varepsilon_t$ below are i.i.d. $N(0, 1)$.

Block 1

This block is to study the size and power of the tests for functional form or varying coefficients in time-series models. Let $x_t = y_{t-1}$. The following two data generating processes (DGP) are taken from Lee et al. (1993).

DGP 1 Linear AR (1)

$$y_t = 0.6y_{t-1} + \varepsilon_t$$

DGP 2 Threshold Autoregressive [TAR(1)] (Nonlinear AR)

$$y_t = 0.9y_{t-1} + \varepsilon_t \text{ if } |y_{t-1}| \leq 1$$
$$= -0.3y_{t-1} + \varepsilon_t \text{ if } |y_{t-1}| > 1$$

Note that DGP 1 is a constant parameter model whereas the alternative DGP 2 is a varying parameter model $y_t = y_{t-1}\delta(y_{t-1}) + \varepsilon_t$, where $\delta(y_{t-1}) = 0.9$ or $-0.3$, depending on $x_t = z_t = y_{t-1}$. In this sense, testing for DGP 1 against DGP 2 is also a test for varying parameters.

Block 2

This block is to study the size and power of the tests for functional form in cross-sectional models. Let $v_{t1}$ and $v_{t2}$ be drawn from $N(0, 1)$. Two regressors $x_{t1}$ and $x_{t2}$ are defined as $x_{t1} = v_{t1}$ and $x_{t2} = (v_{t1} + v_{t2})/\sqrt{2}$. Let $x_t = (x_{t1} x_{t2})$. The following two models are taken from Zheng (1996).

DGP 3

$$y_t = 1 + x_{t1} + x_{t2} + \varepsilon_t$$
DGP 4

\[ y_t = |1 + x_{t1} + x_{t2}|^{5/3} + \varepsilon_t \]

Block 3

This block is to study the size and power of the tests for lag selection in time-series models. Let \( x_t = (y_{t-1} \ y_{t-2}) \). The alternative model DGP 6 is taken from Cai et al. (2000).

DGP 5  Exponential AR(1)

\[ y_t = a_1(y_{t-1})y_{t-1} + 0.2\varepsilon_t \]
\[ a_1(y_{t-1}) = 0.138 + (0.316 + 0.982y_{t-1})\exp(-3.89y_{t-1}^2) \]

DGP 6  Exponential AR(2)

\[ y_t = a_1(y_{t-1})y_{t-1} + a_2(y_{t-1})y_{t-2} + 0.2\varepsilon_t \]
\[ a_1(y_{t-1}) = 0.138 + (0.316 + 0.982y_{t-1})\exp(-3.89y_{t-1}^2) \]
\[ a_2(y_{t-1}) = -0.437 - (0.659 + 1.260y_{t-1})\exp(-3.89y_{t-1}^2) \]

Block 4

This block is to study the size and power of the tests for variable selection in cross-sectional models. Let \( \nu_{t1} \) and \( \nu_{t2} \) be drawn from \( \text{IN}(0,1) \). Two regressors \( x_{t1} \) and \( x_{t2} \) are defined as \( x_{t1} = \nu_{t1} \) and \( x_{t2} = (\nu_{t1} + \nu_{t2})/\sqrt{2} \). Let \( x_t = (x_{t1} \ x_{t2}) \). The alternative model DGP 8 is taken from Zheng (1996).

DGP 7

\[ y_t = |1 + x_{t1}|^{5/3} + \varepsilon_t \]

DGP 8

\[ y_t = |1 + x_{t1} + x_{t2}|^{5/3} + \varepsilon_t \]

To estimate \( \hat{u}_t \) for the null model and \( \tilde{u}_t \) for the alternative model, the information sets used are \( x_t = y_{t-1} \) for Block 1, \( x_t = (y_{t-1} \ y_{t-2}) \) for Block 3, and \( x_t = (x_{t1} \ x_{t2}) \) for Blocks 2 and 4. The omitted variable is \( y_{t-2} \) for Block 3 and \( x_{t2} \) for Block 4.

We use a scalar “threshold variable” \( z_t \) for all models: \( z_t = y_{t-1} \) for Blocks 1 and 3, and \( z_t = x_{t1} \) for Blocks 2 and 4.
For the $Q_a$, $R_a$, $S_a$, and $T_a$ tests, as suggested by Cai et al. (2000), we select $h$ using out-of-sample cross-validation. Let $m$ and $Q$ be two positive integers such that $n > mQ$. The basic idea is first to use $Q$ subseries of lengths $n/qm$ ($q = 1, \ldots, Q$) to estimate the coefficient functions $\delta_q(z)$ and then to compute the one-step forecast errors of the next segment of the time series of length $m$ based on the estimated models. That is, to choose $h$ minimizing the average of the mean square forecast errors:

$$AMS(h) = \frac{1}{m} \sum_{t=n-qm+1}^{n} [y_t - X_t \tilde{\delta}_q(z)]^2$$

and $\tilde{\delta}_q(z)$ are computed from the sample \{$y_t, x_t\}_{t=1}^{n-qm}$. We use $m = [0.1n]$, $Q = 4$, and the Epanechnikov kernel $K(z) = (3/4)(1-z^2)1(|z| < 1)$.

For the $L$ test, as in Li and Wang (1998, p. 154), we use a standard normal kernel. Note that $x_t$ is an $1 \times k$ vector, and $k = 1$ for Block 1 and $k = 2$ for Blocks 2, 3, and 4. Thus, the smoothing parameter $h$ is chosen as $h_i = c\hat{\sigma}_i n^{-1/5}$ ($i = 1$) for the cases with $k = 1$, and $h_i = c\hat{\sigma}_i n^{-1/6}$ ($i = 1, 2$) for the cases with $k = 2$, where $\hat{\sigma}_i$ is the sample standard deviation of $i$th element of $x_t$. The four values of $c = 0.1, 0.5, 1,$ and 2 are used, and the corresponding estimated rejection probability will be denoted as $L_c$.

Test statistics are denoted as $Q^j_a$, $R^j_a$, $S^j_a$, $T^j_a$, and $L^j_c$, with the superscripts $j = A, B, W$ referring to the methods of obtaining the null distributions of the test statistics; asymptotics ($j = A$), naive bootstrap ($j = B$), and wild bootstrap ($j = W$).

Monte Carlo experiments are conducted with 500 bootstrap resamples and 1000 Monte Carlo replications. The amount of computing time needed to obtain the size results (Panels A and B) of the 36 statistics shown in the tables for both 5 and 10% levels was as follows. It took a 600 MHz–256 MB Pentium III PC approximately 1 day for $n = 50$, 2–3 days for $n = 100$, and 5–7 days for $n = 200$. So, it took the PC roughly 7–10 days for the size results in each table. It took another 7–10 days for the power results (Panels C and D) in each table. It took less for Tables 1 and 2 where the null hypothesis is Eq. (11) than for Tables 3 and 4 where the null hypothesis is Eq. (12) and thus both the null and alternative models are nonparametric. For all the results in this chapter, it took the PC about 2 months. A GAUSS code for computing all the tests is available from the authors.
### Table 1: Testing for Linearity in Time-Series Models (Block 1)

<table>
<thead>
<tr>
<th>n</th>
<th>$Q^\beta_\infty$</th>
<th>$Q^W_\infty$</th>
<th>$Q^\beta_1$</th>
<th>$Q^W_1$</th>
<th>$Q^\beta_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$R^\beta_\infty$</th>
<th>$R^W_\infty$</th>
<th>$R^\beta_1$</th>
<th>$R^W_1$</th>
<th>$R^\beta_{15}$</th>
<th>$R^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.089 0.125 0.089 0.099 0.081 0.075 0.102 0.135 0.101 0.117 0.097 0.089</td>
<td>100  0.041 0.072 0.029 0.036 0.030 0.024 0.049 0.087 0.039 0.050 0.036 0.036</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.033 0.059 0.024 0.026 0.021 0.019 0.037 0.066 0.033 0.036 0.027 0.024</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$S^\beta_\infty$</th>
<th>$S^W_\infty$</th>
<th>$S^\beta_1$</th>
<th>$S^W_1$</th>
<th>$S^\beta_{15}$</th>
<th>$S^W_{15}$</th>
<th>$T^\beta_\infty$</th>
<th>$T^W_\infty$</th>
<th>$T^\beta_1$</th>
<th>$T^W_1$</th>
<th>$T^\beta_{15}$</th>
<th>$T^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.022 0.092 0.025 0.073 0.020 0.055 0.022 0.092 0.025 0.073 0.020 0.055 0.065</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.018 0.062 0.015 0.031 0.019 0.020 0.018 0.062 0.015 0.031 0.019 0.020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.021 0.057 0.024 0.028 0.015 0.018 0.021 0.057 0.024 0.028 0.015 0.018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$L^\beta_{0,1}$</th>
<th>$L^W_{0,1}$</th>
<th>$L^\beta_{0,5}$</th>
<th>$L^W_{0,5}$</th>
<th>$L^\beta_1$</th>
<th>$L^W_1$</th>
<th>$L^\beta_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\beta_1$</th>
<th>$L^W_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.047 0.049 0.048 0.001 0.046 0.042 0.000 0.039 0.039</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.031 0.041 0.044 0.006 0.049 0.032 0.000 0.048 0.047</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.038 0.045 0.044 0.027 0.052 0.047 0.002 0.050 0.041</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$Q^\beta_\infty$</th>
<th>$Q^W_\infty$</th>
<th>$Q^\beta_1$</th>
<th>$Q^W_1$</th>
<th>$Q^\beta_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$R^\beta_\infty$</th>
<th>$R^W_\infty$</th>
<th>$R^\beta_1$</th>
<th>$R^W_1$</th>
<th>$R^\beta_{15}$</th>
<th>$R^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.109 0.184 0.112 0.127 0.104 0.096 0.126 0.170 0.128 0.147 0.122 0.113</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.057 0.102 0.049 0.051 0.041 0.039 0.080 0.113 0.064 0.069 0.067 0.061</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.048 0.097 0.038 0.036 0.033 0.031 0.064 0.121 0.064 0.056 0.050 0.045</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$S^\beta_\infty$</th>
<th>$S^W_\infty$</th>
<th>$S^\beta_1$</th>
<th>$S^W_1$</th>
<th>$S^\beta_{15}$</th>
<th>$S^W_{15}$</th>
<th>$T^\beta_\infty$</th>
<th>$T^W_\infty$</th>
<th>$T^\beta_1$</th>
<th>$T^W_1$</th>
<th>$T^\beta_{15}$</th>
<th>$T^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.048 0.127 0.048 0.105 0.046 0.079 0.048 0.127 0.048 0.105 0.046 0.079</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.041 0.093 0.032 0.048 0.036 0.043 0.041 0.093 0.032 0.048 0.036 0.043</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.048 0.102 0.041 0.044 0.033 0.037 0.048 0.102 0.041 0.044 0.033 0.037</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$L^\beta_{0,1}$</th>
<th>$L^W_{0,1}$</th>
<th>$L^\beta_{0,5}$</th>
<th>$L^W_{0,5}$</th>
<th>$L^\beta_1$</th>
<th>$L^W_1$</th>
<th>$L^\beta_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\beta_1$</th>
<th>$L^W_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.089 0.121 0.117 0.033 0.103 0.094 0.005 0.102 0.101 0.000 0.077 0.094</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.060 0.086 0.086 0.029 0.093 0.063 0.011 0.092 0.096 0.000 0.090 0.097</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.065 0.092 0.103 0.050 0.098 0.085 0.020 0.103 0.100 0.003 0.094 0.090</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. Power of tests at 5% level with DGP 2

<table>
<thead>
<tr>
<th>n</th>
<th>$Q^\beta_\infty$</th>
<th>$Q^W_\infty$</th>
<th>$Q^\beta_1$</th>
<th>$Q^W_1$</th>
<th>$Q^\beta_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$R^\beta_\infty$</th>
<th>$R^W_\infty$</th>
<th>$R^\beta_1$</th>
<th>$R^W_1$</th>
<th>$R^\beta_{15}$</th>
<th>$R^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.422 0.506 0.422 0.452 0.432 0.414 0.457 0.533 0.473 0.488 0.489 0.461</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.649 0.731 0.677 0.674 0.684 0.663 0.717 0.776 0.732 0.741 0.744 0.727</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.922 0.946 0.908 0.907 0.906 0.905 0.954 0.965 0.947 0.948 0.948 0.956</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Table 1 presents the empirical size (DGP 1) and power (DGP 2) of testing for neglected nonlinearity in time-series models in Block 1. We observe the following:

1. The $L$ test using bootstrap ($L^B$ and $L^W$) exhibits excellent size behavior and is better than all the other tests ($L^A$, $Q^{j}$, $R^{j}$, $S^{j}$, and $T^{j}$, $j = A, B, W$).
### Table 2  Testing for Linearity in Cross-Sectional Models (Block 2)

#### A. Size of tests at 5% level with DGP 3

<table>
<thead>
<tr>
<th>n</th>
<th>$Q^b_0$</th>
<th>$Q^W_0$</th>
<th>$Q^b_2$</th>
<th>$Q^W_2$</th>
<th>$Q^b_{1,5}$</th>
<th>$Q^W_{1,5}$</th>
<th>$R^b_\infty$</th>
<th>$R^W_\infty$</th>
<th>$R^b_2$</th>
<th>$R^W_2$</th>
<th>$R^b_{1,5}$</th>
<th>$R^W_{1,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.109</td>
<td>0.054</td>
<td>0.071</td>
<td>0.040</td>
<td>0.045</td>
<td>0.122</td>
<td>0.061</td>
<td>0.082</td>
<td>0.047</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.028</td>
<td>0.080</td>
<td>0.020</td>
<td>0.033</td>
<td>0.010</td>
<td>0.013</td>
<td>0.040</td>
<td>0.099</td>
<td>0.035</td>
<td>0.043</td>
<td>0.026</td>
<td>0.028</td>
</tr>
<tr>
<td>200</td>
<td>0.020</td>
<td>0.054</td>
<td>0.007</td>
<td>0.012</td>
<td>0.003</td>
<td>0.002</td>
<td>0.024</td>
<td>0.068</td>
<td>0.014</td>
<td>0.018</td>
<td>0.009</td>
<td>0.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$S^b_0$</th>
<th>$S^W_0$</th>
<th>$S^b_2$</th>
<th>$S^W_2$</th>
<th>$S^b_{1,5}$</th>
<th>$S^W_{1,5}$</th>
<th>$T^b_\infty$</th>
<th>$T^W_\infty$</th>
<th>$T^b_2$</th>
<th>$T^W_2$</th>
<th>$T^b_{1,5}$</th>
<th>$T^W_{1,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.014</td>
<td>0.078</td>
<td>0.016</td>
<td>0.052</td>
<td>0.013</td>
<td>0.027</td>
<td>0.014</td>
<td>0.078</td>
<td>0.016</td>
<td>0.052</td>
<td>0.013</td>
<td>0.027</td>
</tr>
<tr>
<td>100</td>
<td>0.015</td>
<td>0.061</td>
<td>0.010</td>
<td>0.021</td>
<td>0.007</td>
<td>0.011</td>
<td>0.015</td>
<td>0.061</td>
<td>0.010</td>
<td>0.021</td>
<td>0.007</td>
<td>0.011</td>
</tr>
<tr>
<td>200</td>
<td>0.012</td>
<td>0.052</td>
<td>0.005</td>
<td>0.014</td>
<td>0.006</td>
<td>0.007</td>
<td>0.012</td>
<td>0.052</td>
<td>0.005</td>
<td>0.014</td>
<td>0.006</td>
<td>0.007</td>
</tr>
</tbody>
</table>

#### B. Size of tests at 10% level with DGP 3

<table>
<thead>
<tr>
<th>n</th>
<th>$L^b_{0,1}$</th>
<th>$L^W_{0,1}$</th>
<th>$L^b_{0,5}$</th>
<th>$L^W_{0,5}$</th>
<th>$L^b_{1}$</th>
<th>$L^W_{1}$</th>
<th>$L^b_{2}$</th>
<th>$L^W_{2}$</th>
<th>$L^b_{1,5}$</th>
<th>$L^W_{1,5}$</th>
<th>$L^b_{2,5}$</th>
<th>$L^W_{2,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.026</td>
<td>0.050</td>
<td>0.030</td>
<td>0.022</td>
<td>0.048</td>
<td>0.049</td>
<td>0.007</td>
<td>0.041</td>
<td>0.000</td>
<td>0.048</td>
<td>0.045</td>
<td>0.048</td>
</tr>
<tr>
<td>100</td>
<td>0.038</td>
<td>0.042</td>
<td>0.038</td>
<td>0.016</td>
<td>0.052</td>
<td>0.061</td>
<td>0.009</td>
<td>0.054</td>
<td>0.001</td>
<td>0.054</td>
<td>0.047</td>
<td>0.049</td>
</tr>
<tr>
<td>200</td>
<td>0.040</td>
<td>0.046</td>
<td>0.054</td>
<td>0.022</td>
<td>0.040</td>
<td>0.042</td>
<td>0.015</td>
<td>0.053</td>
<td>0.001</td>
<td>0.057</td>
<td>0.053</td>
<td>0.062</td>
</tr>
</tbody>
</table>

#### C. Power of tests at 5% level with DGP 4

<table>
<thead>
<tr>
<th>n</th>
<th>$Q^b_0$</th>
<th>$Q^W_0$</th>
<th>$Q^b_2$</th>
<th>$Q^W_2$</th>
<th>$Q^b_{1,5}$</th>
<th>$Q^W_{1,5}$</th>
<th>$R^b_\infty$</th>
<th>$R^W_\infty$</th>
<th>$R^b_2$</th>
<th>$R^W_2$</th>
<th>$R^b_{1,5}$</th>
<th>$R^W_{1,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(continued)
2. R is better than S, and T. S, and T are identical; Q behaves similarly to R. The tests of Q, R, S, T tend to be oversized for n = 50, and undersized with n = 100 or 200, which is more apparent with the larger sample size.

3. Trimming for \(R_a\), \(S_a\), \(T_a\), and \(Q_a\) is useful when n is small. For example, for n = 50, \(T_2\) works better than \(T_\infty\). However, the trimming makes the size worse when n is large (say, n = 200).

4. The asymptotic test \(L^A\) works better with smaller \(c\). \(L^A\) is not reliable for \(c > 0.1\) and gets worse as \(c\) gets larger. The bootstrap tests \(L^B\) and \(L^W\) work very well with all four values of \(c\). The asymptotic \(L^A\) is sensitive to \(c\) while the bootstrap tests \(L^B\) and \(L^W\) are not sensitive to \(c\). This is different from what is found by Lee and Ullah (2001) where the bootstrap tests \(L^B\)
### Table 3  Lag Selection in Time-Series Models (Block 3)

#### A. Size of tests at 5% level with DGP 5

<table>
<thead>
<tr>
<th>n</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q^B_∞</td>
<td>0.073</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td>Q^W_∞</td>
<td>0.106</td>
<td>0.034</td>
<td>0.025</td>
</tr>
<tr>
<td>Q^B_2</td>
<td>0.099</td>
<td>0.015</td>
<td>0.012</td>
</tr>
<tr>
<td>Q^W_2</td>
<td>0.087</td>
<td>0.022</td>
<td>0.016</td>
</tr>
<tr>
<td>Q^B_15</td>
<td>0.094</td>
<td>0.019</td>
<td>0.014</td>
</tr>
<tr>
<td>Q^W_15</td>
<td>0.078</td>
<td>0.013</td>
<td>0.011</td>
</tr>
<tr>
<td>R^B_∞</td>
<td>0.084</td>
<td>0.037</td>
<td>0.028</td>
</tr>
<tr>
<td>R^W_∞</td>
<td>0.099</td>
<td>0.019</td>
<td>0.016</td>
</tr>
<tr>
<td>R^B_2</td>
<td>0.088</td>
<td>0.023</td>
<td>0.020</td>
</tr>
<tr>
<td>R^W_2</td>
<td>0.096</td>
<td>0.021</td>
<td>0.017</td>
</tr>
</tbody>
</table>

#### B. Size of tests at 10% level with DGP 5

<table>
<thead>
<tr>
<th>n</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q^B_∞</td>
<td>0.085</td>
<td>0.022</td>
<td>0.002</td>
</tr>
<tr>
<td>Q^W_∞</td>
<td>0.124</td>
<td>0.051</td>
<td>0.021</td>
</tr>
<tr>
<td>Q^B_2</td>
<td>0.098</td>
<td>0.030</td>
<td>0.010</td>
</tr>
<tr>
<td>Q^W_2</td>
<td>0.121</td>
<td>0.044</td>
<td>0.014</td>
</tr>
<tr>
<td>Q^B_15</td>
<td>0.119</td>
<td>0.031</td>
<td>0.011</td>
</tr>
<tr>
<td>Q^W_15</td>
<td>0.060</td>
<td>0.023</td>
<td>0.009</td>
</tr>
<tr>
<td>R^B_∞</td>
<td>0.100</td>
<td>0.023</td>
<td>0.002</td>
</tr>
<tr>
<td>R^W_∞</td>
<td>0.121</td>
<td>0.061</td>
<td>0.009</td>
</tr>
<tr>
<td>R^B_2</td>
<td>0.108</td>
<td>0.043</td>
<td>0.011</td>
</tr>
<tr>
<td>R^W_2</td>
<td>0.114</td>
<td>0.035</td>
<td>0.009</td>
</tr>
</tbody>
</table>

#### C. Power of tests at 5% level with DGP 6

<table>
<thead>
<tr>
<th>n</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q^B_∞</td>
<td>0.983</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Q^W_∞</td>
<td>0.988</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Q^B_2</td>
<td>0.988</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Q^W_2</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Q^B_15</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Q^W_15</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(continued)
and $L^W$ are also sensitive to $c$. This is because $\{y^*_t\}$ in this chapter is not recursively generated (as described above) while Lee and Ullah (2001) generated $\{y^*_t\}$ recursively for time-series data. Note that in this chapter we generated the bootstrap data $\{y^*_t\}$ conditional on $\{x_t\}$ for both the cases when $x_t$ are exogenous (Blocks 2 and 4) and the cases when $x_t$ are lagged dependent variables (Blocks 1 and 3). The bootstrap method used in Lee and Ullah (2001) may be called the “recursive” bootstrap, while the bootstrap method used in this chapter may be called the “conditional” bootstrap. As discussed in Lee (2001), the bootstrap method treating $x_t$ as given and generating $\{y^*_t\}$ conditional on $x_t$ gives more robust size behavior than the recursive bootstrap even for the time-series data.

### Table 3

Continued

<table>
<thead>
<tr>
<th>$n$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.995</td>
<td>0.998</td>
<td>0.995</td>
<td>0.998</td>
<td>0.997</td>
<td>0.998</td>
<td>0.995</td>
<td>0.998</td>
<td>0.995</td>
<td>0.998</td>
<td>0.995</td>
<td>0.998</td>
<td>0.995</td>
<td>0.998</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.282</td>
<td>0.496</td>
<td>0.336</td>
<td>0.965</td>
<td>0.976</td>
<td>0.979</td>
<td>0.985</td>
<td>0.982</td>
<td>0.980</td>
<td>0.986</td>
<td>0.994</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.828</td>
<td>0.899</td>
<td>0.836</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D. Power of tests at 10% level with DGP 6

<table>
<thead>
<tr>
<th>$n$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
<th>$Q^\phi_{15}$</th>
<th>$Q^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.986</td>
<td>0.990</td>
<td>0.987</td>
<td>0.991</td>
<td>0.991</td>
<td>0.991</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.998</td>
<td>1.000</td>
<td>0.998</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
<th>$S^\phi_{15}$</th>
<th>$S^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.997</td>
<td>0.998</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
<th>$L^\phi_{15}$</th>
<th>$L^W_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.571</td>
<td>0.680</td>
<td>0.550</td>
<td>0.979</td>
<td>0.978</td>
<td>0.978</td>
<td>0.985</td>
<td>0.990</td>
<td>0.984</td>
<td>0.991</td>
<td>0.996</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.946</td>
<td>0.961</td>
<td>0.931</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
### Table 4 Variable selection in Cross-Sectional Models (Block 4)

#### A. Size of tests at 5% level with DGP 7

<table>
<thead>
<tr>
<th>n</th>
<th>$Q_\infty^B$</th>
<th>$Q_\infty^W$</th>
<th>$Q_2^B$</th>
<th>$Q_2^W$</th>
<th>$Q_{15}^B$</th>
<th>$Q_{15}^W$</th>
<th>$R_\infty^B$</th>
<th>$R_\infty^W$</th>
<th>$R_2^B$</th>
<th>$R_2^W$</th>
<th>$R_{15}^B$</th>
<th>$R_{15}^W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.025</td>
<td>0.055</td>
<td>0.032</td>
<td>0.043</td>
<td>0.032</td>
<td>0.036</td>
<td>0.029</td>
<td>0.048</td>
<td>0.035</td>
<td>0.042</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td>100</td>
<td>0.011</td>
<td>0.022</td>
<td>0.014</td>
<td>0.021</td>
<td>0.011</td>
<td>0.013</td>
<td>0.013</td>
<td>0.028</td>
<td>0.019</td>
<td>0.024</td>
<td>0.021</td>
<td>0.018</td>
</tr>
<tr>
<td>200</td>
<td>0.003</td>
<td>0.014</td>
<td>0.011</td>
<td>0.015</td>
<td>0.008</td>
<td>0.010</td>
<td>0.004</td>
<td>0.013</td>
<td>0.007</td>
<td>0.014</td>
<td>0.009</td>
<td>0.009</td>
</tr>
</tbody>
</table>

#### B. Size of tests at 10% level with DGP 7

<table>
<thead>
<tr>
<th>n</th>
<th>$S_\infty^B$</th>
<th>$S_\infty^W$</th>
<th>$S_2^B$</th>
<th>$S_2^W$</th>
<th>$S_{15}^B$</th>
<th>$S_{15}^W$</th>
<th>$T_\infty^B$</th>
<th>$T_\infty^W$</th>
<th>$T_2^B$</th>
<th>$T_2^W$</th>
<th>$T_{15}^B$</th>
<th>$T_{15}^W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.030</td>
<td>0.024</td>
<td>0.007</td>
<td>0.020</td>
<td>0.010</td>
<td>0.022</td>
<td>0.003</td>
<td>0.024</td>
<td>0.007</td>
<td>0.020</td>
<td>0.010</td>
<td>0.022</td>
</tr>
<tr>
<td>100</td>
<td>0.006</td>
<td>0.016</td>
<td>0.007</td>
<td>0.013</td>
<td>0.008</td>
<td>0.011</td>
<td>0.006</td>
<td>0.017</td>
<td>0.007</td>
<td>0.013</td>
<td>0.008</td>
<td>0.011</td>
</tr>
<tr>
<td>200</td>
<td>0.002</td>
<td>0.010</td>
<td>0.005</td>
<td>0.011</td>
<td>0.005</td>
<td>0.007</td>
<td>0.002</td>
<td>0.010</td>
<td>0.005</td>
<td>0.011</td>
<td>0.005</td>
<td>0.007</td>
</tr>
</tbody>
</table>

#### C. Power of tests at 5% level with DGP 8

<table>
<thead>
<tr>
<th>n</th>
<th>$Q_\infty^B$</th>
<th>$Q_\infty^W$</th>
<th>$Q_2^B$</th>
<th>$Q_2^W$</th>
<th>$Q_{15}^B$</th>
<th>$Q_{15}^W$</th>
<th>$R_\infty^B$</th>
<th>$R_\infty^W$</th>
<th>$R_2^B$</th>
<th>$R_2^W$</th>
<th>$R_{15}^B$</th>
<th>$R_{15}^W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.993</td>
<td>0.993</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
<td>0.995</td>
<td>0.995</td>
<td>0.994</td>
<td>0.994</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(continued)
5. Turning to the power behavior, although the size of $L_B$ and $L_W$ are quite robust to $c$, the power of these tests can vary with $c$ and is generally best with larger $c$. The tests $Q$, $R$, $S$, $T$ have a similar power pattern but these are slightly worse than $L$.

Table 2 presents the size (DGP 3) and power (DGP 4) of testing for neglected nonlinearity in cross-sectional models in Block 2. The following observations are made. All the size results in Table 1 for time series (summarized above) hold here for Table 2 with cross-sectional data. While the size of $L_B$ and $L_W$ are quite robust to $c$, the power of these tests can vary with $c$ and is higher with larger $c$.

Table 3 presents the size (DGP 5) and power (DGP 6) of testing for lag selection in time-series models in Block 3. The results are very similar
to those in Tables 1 and 2. The tests $L^B$ and $L^W$ have good size and power in testing for lag selection in nonparametric time-series models as well as in testing for parametric functional forms.

Table 4 presents the size (DGP 7) and power (DGP 8) of testing for variable selection in cross-sectional models in Block 4. The results are again very similar to those in Tables 1–3. The tests $L^B$ and $L^W$ have very good size and power in testing for omitted variables in cross-sectional models.

5 CONCLUSIONS

We consider three nonparametric tests for functional form, varying parameters, and omitted variables in regression models both of time-series data and of cross-sectional data. The first approach ($R, S, T$) is to compare the sums of squared residuals from the null and the alternative models and the second test ($Q$) is to compare the fitted values of the null and alternative models. The third test ($L$) is the nonparametric conditional moment test, which is to see if the residuals from the null model are related to the conditioning variables in the alternative models. We find that the bootstrap tests of Li and Wang (1998) and Zheng (1996), $L^B$ and $L^W$, have very good size and power properties in all the situations we considered.

One of the reasons for the better performance of these $L$ tests compared to $R, S, T$, and $Q$ tests may be because the asymptotic distribution of $L$ is asymptotically normal with the mean zero under the null hypothesis, whereas this is not the case for $R, S$, and $T$ tests. Therefore, it will be an interesting future study to compare the $L$ test with the bias-adjusted $R, S, T$ tests as described in Fan and Li (2001). It will also be useful to develop the theoretical power properties of the tests under various local alternatives, as studied in Hong and Lee (2001) and Tripathi and Kitamura (2000) in different but related contexts. Moreover, this chapter has considered the tests based on the kernel smoothing procedure only. It will be useful to study how our study compares with other specification testing procedures, especially using other smoothing procedures such as neural network, spline regression, and fuzzy c-means algorithm in Giles and Draeseke (this volume, Chap. 14). Finally, the issue of optimal choice of window-width for the tests considered here needs further future investigations.

ACKNOWLEDGMENTS

We thank Zongwu Cai, Jianqing Fan, and Qi Li for their programs and comments, and the UCR Academic Senate for their research support.
REFERENCES


Lewbel, A., 1993. Consistent tests with nonparametric components with an application to Chinese production data. Unpublished manuscript, Brandeis University, Waltham, MA.


Li, Q., Huang, C.J., and Fu, T-T., 1997. Semiparametric smooth coefficient models. Unpublished manuscript, Texas A&M University, College Station, TX.


Nonparametric Bootstrap Specification Testing


The Effect of Economic Growth on Standard of Living: A Semiparametric Analysis

Nilanjana Roy
University of Victoria, Victoria, British Columbia, Canada

1 INTRODUCTION

Up until the late 1960s, development was synonymous with economic growth, the idea being that the increase in national income would eventually “trickle down” to the masses and the standard of living of the people would improve. However, the experience of the developing countries in the 1970s did not lend empirical support to this view. So the view started changing in the 1970s and the idea of development became more people oriented than goods oriented. Development policies started focusing directly on the well-being of the people or on raising the standard of living of the people rather than on increasing the national income of the country.

The question for development economists and policymakers then became: “How to improve the standard of living of the people?” and a related question that was raised was “Does economic growth positively affect the standard of living, and if so, how strong is the influence?” The latter is an important question since an affirmative answer to that question would provide support for policies that are aimed at increasing income in developing countries. However, the answer to the latter question was not
easy to find and, in fact, the debate still goes on. According to some researchers (Iseman, 1980; Sen, 1981; Anand and Kanbur, 1991; Anand and Ravallion, 1993; among others), it is not income but public provision of services that plays an important role in raising the standard of living. While others, such as Bhalla (1988), Kakwani (1993), Pritchett and Summers (1996), and Chakraborty (1999)* have concluded from their analyses that there exists a positive and statistically significant relationship between \textit{per capita} income and standard of living. It should be noted here, that standard of living has been measured by various indicators in the literature. Some researchers have used life expectancy at birth, others have used infant mortality, still others have used composite indexes combining health indicators and indicators for literacy, among others.

One of the contributions of Pritchett and Summers (1996) is that they controlled for unobserved country-specific effects such as good government or cultural values that could affect the standard of living indicator positively. However, they did not pay much attention to the issue of functional form specifications and stated “We didn’t consider the issue of functional form in the context of the present work to be sufficiently important to merit formal specification tests”. However, it is well known in the econometrics literature that mis-specification of functional form can lead to inconsistent and inefficient estimates and suboptimal tests.

In fact, the nonlinearity in the relationship between \textit{per capita} income and various indicators of standard of living has been recognized by many researchers and some of the functional forms used in this literature are logistic, log-linear, and semilog, among others. However, as Chakraborty (1999) correctly points out, no formal specification testing was done in any of these studies and the various specifications were chosen arbitrarily. Chakraborty (1999) contributes to the literature by using an alternative methodology, namely, the nonparametric approach, to study the relation between standard of living and \textit{per capita} income. This data-dependent approach does not need the specification of any functional form and hence avoid the problem of functional form mis-specification.

However, the rate of convergence of nonparametric estimators are slower than that of their parametric counterparts. So it is meaningful to do specification testing to see if the data supports the parametric model, before we move to the more computationally involved nonparametric estimator. Chakraborty (1999) does not undertake any such specification testing to provide support for her preferred nonparametric model. Given

*This paper has a very nice survey of the empirical literature.

that one of the aims of this chapter is to undertake such specification testing to see if the use of the nonparametric models are appropriate in studying the relationship between standard of living and \textit{per capita} income, we now briefly discuss some of the issues dealt with specifically in Chakraborty (1999).

Chakraborty (1999) estimated a nonparametric regression of the “achievement function” on the logarithm of \textit{per capita} income and a nonparametric regression of the “improvement function” on the logarithm of \textit{per capita} income and its change. The concepts of the “achievement function” and the “improvement function” were developed by Kakwani (1993) using an axiomatic approach. The achievement function is a transformation of some standard of living indicator such as life expectancy and was constructed such that it gives different weights to equal increments in the indicator, depending on the initial level of the indicator. More specifically, the function gives higher weights to those countries that have already achieved a higher level of longevity, say, since it is relatively more difficult to increase longevity from a higher level than from a lower level. The achievement function is given as

\[
f(x, M_0, M) = \frac{\ln(M - M_0) - \ln(M - x)}{\ln(M - M_0)}
\]

where \( x \) denotes the value of the positive standard of living indicator for the country in question, e.g., it can be the country’s life expectancy, while \( M \) and \( M_0 \) denote the upper and lower limits of the indicator. For life expectancy, the limits are generally set at 80 and 30 respectively. Also, note that for an indicator such as infant mortality, which is a negative measure of the standard of living, the above formula needs to be adjusted as

\[
f(I, m_0, m) = \frac{\ln(m - m_0) - \ln(I - m_0)}{\ln(m - m_0)}
\]

where \( I \) is the infant mortality rate and \( m \) and \( m_0 \) are its upper and lower limits, set generally at 300 and 5, respectively. See Kakwani (1993) for details on the achievement function. The improvement functions corresponding to Eqs. (1) and (2) are defined as the difference in the achievement function in two given periods. According to Kakwani (1993), the nonlinearity in the relationship between a standard of living indicator and the \textit{per capita} income is captured adequately by specifying a linear relationship between the achievement function and the logarithm of \textit{per capita} income. Hence, he estimated linear models using achievement function as the dependent variable and logarithm of \textit{per capita} income as the independent variable. Chakraborty (1999) claimed that this model was...
mis-specified and used a nonparametric specification instead. She also detected nonlinearity in the relationship between the improvement function and the level of *per capita* income and its difference and again proposed using a purely nonparametric specification. The results from her specifications were contradictory to Kakwani’s (1993) result that a country’s level of economic welfare has a positive and statistically significant impact on its improvement in the living standard. In this chapter, using the same data set, we formally test Kakwani’s linear specification against Chakraborty’s nonparametric specification using Li and Wang’s test (1998).

Neither Kakwani (1993) nor Chakraborty (1999) considered the effect of including public health expenditure per head as a regressor in the model. The second aim of this chapter is to include the new variable in the regression to examine its effect on the estimated coefficient of the income variable. As we discussed before, some researchers claim that it is public provision of services that plays an important role in raising the living standard and the income coefficient will lose its strength and statistical significance once the other variable has been accounted for (Anand and Ravallion, 1993). We have tested a parametric specification against a semiparametric one using Li and Wang’s test (1998) and based on the results from the test, we have presented a semiparametric regression model with the logarithm of the public health expenditure per head appearing nonlinearly in the model. Our results show that irrespective of the specification, the income coefficient continues to be statistically significant. The semiparametric specification allows us to provide a $\sqrt{n}$ consistent estimator of the income coefficient and also allows us to offer a graphical representation of the nonlinear behavior of the logarithm of the health expenditure variable.

Section 2.1 offers a brief discussion of the semiparametric estimation method (Robinson, 1988), Section 2.2 provides an outline of the specification test of Li and Wang (1998), while Section 2.3 contains a discussion of some of the computational issues related to the estimation methods and the test procedures. Section 3.1 gives a brief description of the data while Section 3.2 has the empirical results. Section 4 concludes.

## 2 METHODOLOGY

### 2.1 Semiparametric Model

The semiparametric model can be written as

$$y_i = X_i'\beta + m(Z_i) + u_i \quad i = 1, \ldots, n$$

(3)
where $X_i$ contains $q$ regressors, which enter the specification linearly, $\beta$ is a $q \times 1$ vector of parameters, $Z_i$ contains $p$ continuous variables, and $m(\cdot)$ is an unknown function; $u_i$ is the error term and satisfies the standard assumptions about the error term. Following Robinson (1988), we can obtain a $\sqrt{n}$ consistent estimator of $\beta$ by first taking an expectation of Eq. (3) conditional on $Z_i$ and then subtracting that from Eq. (3) to remove the effect of $m(Z_i)$, and finally estimating the parameter $\hat{\beta}$ in the second step by using the transformed variables where the expectation terms are replaced by their kernel estimates.

We can then rewrite Eq. (3) as

$$y_i - X_i^T \hat{\beta} = m(Z_i) + (u_i + X_i^T (\beta - \hat{\beta}))$$

where the term in the parenthesis on the right-hand side is the new composite error term. Given that in our case $Z_i$ contains only one variable, we can estimate Eq. (4) using standard nonparametric methods to obtain an estimator for $m(Z_i)$. If $Z_i$ contained more than one variable, then to obtain the individual effects of the variables, one can follow the additive partially linear approach and use marginal integration to estimate the components. See Linton and Nielsen (1995) and Fan and Li (1996), among others, for more details on that.

2.2 Specification Test

The two kinds of specification tests that we have used in this chapter have been proposed by Li and Wang (1998). The first one is to test the null of a linear regression model against a purely nonparametric specification. The null hypothesis is given as

$$H_0 : y_i = g(X_i, \beta) + u_i \quad i = 1, \ldots, n$$

and the alternative is given as

$$H_1 : y_i = m(X_i) + v_i \quad i = 1, \ldots, n$$

where $u_i$ and $v_i$ are assumed to have conditional mean zero and conditional variance $\sigma_u^2(X_i)$ and $\sigma_v^2(X_i)$, respectively. For our case, $g(X_i, \beta) = X_i^T \beta$, but the test holds as long as $g(\cdot)$ is some known function.

The test statistic is given by

$$J_n = nh^{1/2} l_n/\sqrt{\Omega}$$  (5)
where \( I_n = (1/n^2 h^q) \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \hat{u}_i \hat{u}_j K_{ij}, \ K_{ij} = K((X_i - X_j)/h), \) is the kernel function, \( h \) is the smoothing parameter, \( \hat{u}_i \) is the \( i \)th least squares residual, and

\[
\hat{\Omega} = \frac{2}{n^2 h^q} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} \hat{u}_i^2 \hat{u}_j^2 K_{ij}^2
\]

Li and Wang (1998) show that the test statistic has an asymptotic standard normal distribution under the null hypothesis. However, in small samples they suggest a wild bootstrap method to obtain the critical values. The rejection region for the test at the level \( \alpha \) is \( J_n > c \) where \( c \) is the critical value. In this chapter, we have chosen the kernel function to be the normal kernel and have reported the bootstrapped critical values.

The second test that we have used is to test the null of a linear regression model augmented by a square term against a semiparametric alternative. The null hypothesis is given as

\[
H_0 : y_i = X'_i \beta + g(Z_i, \gamma) + u_i \quad i = 1, \ldots, n
\]

and the alternative is given as

\[
H_1 : y_i = X'_i \beta + m(Z_i) + v_i \quad i = 1, \ldots, n
\]

where in our particular case, \( g(Z_i, \gamma) = \gamma_1 Z_i + \gamma_2 Z_i^2 \) and the dimension of \( Z_i \) is \( p = 1 \). The test statistic is the same as in Eq. (5) except that the \( q \) is replaced by \( p \) and the \( \hat{u}_i \) is replaced by \( u_{ni} \) where \( u_{ni} = y_i - \hat{X}'_i \hat{\beta} - g(Z_i, \hat{\gamma}) \) is the residual from the “mixed” regression, \( \hat{\beta} \) is any semiparametric \( \sqrt{n} \) consistent estimator of \( \beta \) based on the alternative model,* and \( \hat{\gamma} \) is the least-squares estimator based on the null model. The test statistic is again asymptotically distributed as standard normal but Li and Wang (1998), based on their Monte Carlo Work, suggest the use of bootstrapped critical values.

### 2.3 Computational Issues

The kernel functions used throughout the chapter have been the normal kernel. The bandwidths in this chapter have been chosen using the formula

\[
h = c \times \text{sd} \times n^{-1/2(q+4)},
\]

where “sd” is the estimated standard deviation of the relevant regressor, \( q \) is the number of regressors appearing nonparametrically in the model, and \( c \) is a constant. The results for the different tests have been presented for a range of \( c \) values\(^5\) to show the

---

*See Heckman (1986) and Robinson (1988), among others, for such a choice. We used the Robinson estimator for this chapter.

\(^5\) The \( c \) values used were 0.8, 1.0, 1.2, and 1.4.
robustness of the results to the choice of the bandwidth. The semiparametric estimations were also undertaken for the same range of $c$ values, but the results and figures presented here are for only $c = 1$, partly because visually this bandwidth seemed to offer the correct degree of smoothness and partly due to space constraint. In other words, the smaller bandwidths offered undersmoothing while the larger ones offered oversmoothing. One can also use some computer-intensive approach to calculate the bandwidth, such as the cross-validation method, but we did not use any such method here since the finite sample performances of such approaches are often not very good and our data set has less than 100 observations.

The semiparametric estimations and all the Li and Wang tests were coded and calculated using GAUSS Version 3.2.34. The computation is not time consuming; in fact, none of them required more than 5 min to run. The parametric estimation results and the scatter plots were all calculated using STATA Version 5.0, because of the ease with which panel data can be handled in STATA.

3 EMPIRICAL RESULTS

3.1 Description of the Data

The first part of our analysis uses the data set used by Kakwani (1993). We use the same data set since we want formally to test Chakraborty’s (1999) claim that the parametric models used by Kakwani are mis-specified and that a nonparametric model would be a better alternative even for that particular data set.

The data are based on 80 developing countries and contain information on the population weighted average of the achievement index for each country and the population weighted average GDP per capita in purchasing power parity (PPP) U.S. dollar for the two subperiods, 1971–1980 and 1981–1990. Kakwani used the logarithm of the per capita GDP in PPP U.S. dollar of a country as a measure of its economic welfare. Kakwani (1993) also provides data on the improvement index of each country measured as the difference in the achievement index in the two subperiods. We will focus on the achievement and improvement indices for life expectancy since Chakraborty (1999) focused on that particular variable. For details on the data, see Kakwani (1993, p. 316).

The second part of our analysis extends to a trivariate model similar to that of Anand and Kanbur (1993) but unlike them, we use a semiparametric approach and use data for more than one year. The latter allows us to estimate panel data models and to examine Prichett and
Summers’ (1966, p. 847) observation that good health might be “associated with some other country-level factor such as ‘cultural values’ or ‘good government,’ that is either intrinsically unobservable or for some other reason excluded from the estimation.”

The data used in this part of the analysis have been obtained from the Global Development Network (GDN) growth database.* This database, put together by William Easterly and Hairong Yu, contains panel data for 122 countries on more than 100 variables. The data on the different variables have been collected from many different reliable data sources and the source for each variable is clearly indicated in this database. From the GDN growth database, we collected data on infant mortality rate (per 1000 live births), real gross domestic product per capita in constant dollars (international prices, base year 1985), and health expenditure as percentage of GDP. The original source of the infant mortality data is Global Development Finance and World Development Indicators, that for the real GDP data is the Penn World Table 5.6 and that for the health expenditure data is the IMF: Government Financial Statistics. The data used for our extended model constitute a balanced panel for 26 countries for 3 different years (1982, 1987, and 1992). We have focused our analysis on developing countries only, in keeping with this literature. We chose countries with real per capita GDP in constant dollars less than $6000, as in Pritchett and Summers (1996).

3.2 Results

Kakwani (1993) estimated a linear regression model with the achievement index as the dependent variable and the welfare of a country as the independent variable for each subperiod. He also estimated a linear regression model with the improvement index as the dependent variable and the welfare level and the change in welfare level of a country between the two subperiods as the independent variables. Chakraborty (1999), using a nonparametric framework of analysis and focusing on the achievement index and improvement index related to life expectancy only, concluded that the functional forms of Kakwani’s models were misspecified and detected nonlinearity in the relationship. However,

---

*The database can be found at www.worldbank.org/research/growth/GDNdata.htm

†In fact, there are four countries in our data set that have real per capita GDP greater than $6000 but less than $8000 for the year 1992. We included these countries (Korea, Malta, Mauritius, and Mexico) in our data set since their GDP figures in the other 2 years are less than $6000.
Chakraborty’s conclusions were based on a visual inspection of the nonparametric regression plots, using Kakwani’s data set.*

One of the aims of this study is to test formally for this suspected mis-specification of functional form by using the same data set. Table 1 presents the results from using the nonparametric specification test by Li and Wang (1998), which tests a simple linear model against a purely nonparametric one. The first nine rows of results relate to Kakwani’s bivariate model with achievement index as the dependent variable and the welfare level as the independent variable. The results have been presented for various bandwidths since the choice of bandwidth is usually very important in the nonparametric framework. We have presented bootstrapped critical values since Li and Wang’s (1998) Monte Carlo exercise showed that the wild bootstrap method provided a much better

*It should also be noted that the confidence intervals for the regression curves were not presented.

| Table 1 | Tests of Linear Parametric Models Against Nonparametric Models |
|---------------------------------|
| $c = 0.8$ | $c = 1.0$ | $c = 1.2$ | $c = 1.4$ |
| 1971–1980 (Bivariate model—Kakwani’s data) |
| Test statistic | $-0.406$ | $-0.439$ | $-0.494$ | $-0.549$ |
| c.v. | $0.653$ | $0.478$ | $0.239$ | $0.101$ |
| 1980–1990 (Bivariate model—Kakwani’s data) |
| Test statistic | $0.015$ | $0.055$ | $0.087$ | $0.084$ |
| c.v | $0.839$ | $0.652$ | $0.418$ | $0.273$ |
| 1971–1990 (Bivariate model—Kakwani’s data) |
| Test statistic | $0.843$ | $0.914$ | $0.904$ | $0.847$ |
| c.v | $0.896$ | $0.778$ | $0.483$ | $0.257$ |
| 1971–1990 (Trivariate model—Kakwani’s data) |
| Test statistic | $-0.776$ | $-0.800$ | $-0.870$ | $-0.953$ |
| c.v | $1.239$ | $0.926$ | $0.762$ | $0.622$ |
| 1982, 1987, 1992 (Bivariate model—GDN data, transformed variables) |
| Test statistic | $-0.698$ | $-0.640$ | $-0.589$ | $-0.554$ |
| c.v | $0.899$ | $0.681$ | $0.414$ | $0.292$ |
| 1982, 1987, 1992 (Bivariate model—GDN data, untransformed variables) |
| Test statistic | 2.732 | 2.392 | 2.086 | 1.800 |
| c.v | $-0.390$ | $-0.422$ | $-0.469$ | $-0.493$ |

* c.v. is the bootstrapped 5% critical value based on 399 replications.
null approximation for the test than the asymptotic normality result. For both subperiods, we fail to reject the null hypothesis of the linear model at both the 5 and 10% levels of significance for all the different bandwidths. This result agrees with Li and Wang's (1998) finding that, due to the unbiasedness of the relevant test statistic, this testing procedure is not as sensitive to the choice of \( h \) as many other kernel smoothing procedures. The results lend strong support to Kakwani's model in the bivariate setup and contradict Chakraborty's claim that Kakwani's model relating the achievement index to the welfare level suffered from functional form mis-specification. The test results under the heading 1971–1990 (bivariate model) are for the pooled data set and they show that there is not much support for the null model at the 5% level. However, at the 1% level of significance we fail to reject the model even with the pooled data.* It should be noted though that neither Kakwani (1993) nor Chakraborty (1999) were analysing a pooled model.

The results in Table 1 under the heading 1971–1990 (trivariate model) present the results from Li and Wang’s (1998) test where the null model now is the linear model with the improvement index as the dependent variable and the regressors are the welfare level and the change in welfare level. The alternative model is a purely nonparametric model; i.e., both the regressors appear as unknown functions. The results from the table show that we fail to reject the null model at the 5% level of significance for all \( c \) values considered. In fact, the result holds also at the 10% level of significance. This means that Chakraborty's (1999) claim that this model of Kakwani's suffers from functional form mis-specification is not supported by the data either. So the conclusion from Kakwani (1993, p. 335) that “the level of economic welfare (lagged) and change in it have a positive and significant influence on the improvement in the standard of living” still holds.

The above analysis shows that the use of the achievement index adequately captures the nonlinearity in the relationship between the life expectancy at birth and the welfare level and that there is no further need to model achievement index as an unknown function of the welfare level as suggested by Chakraborty (1999). However, both Kakwani (1993) and Chakraborty (1999) used bivariate models even though both mentioned factors other than income that many have a positive effect on the standard of living over time. One such factor is the public provision of health services as indicated by Anand and Ravallion (1993).

*The figures for the 1% critical value are available from the author on request. In fact, all the results that are mentioned but not shown here are available from the author on request.
Other factors may include unobservable country-specific factors such as good government or cultural values, as indicated by Pritchett and Summers (1996), which would affect health positively. Given this scenario, we use the second data set and examine the role, if any, played by public expenditure on health services and other unobserved country-specific factors in enhancing the health status of a country. We will use infant mortality rate as an indicator of standard of living for this study. The results using life expectancy index can be obtained similarly, but we choose to focus our attention on infant mortality rate mainly since the data on it are more reliable. The life expectancy figures usually are not observed directly from death registrations but are derived by using infant or child mortality figures together with the model life table. This means that they do not contain any more information than the infant or child mortality figures, and in fact suffer from more measurement errors. For further discussion of this issue, see Pritchett and Summers (1996).

Kakwani (1993) and Anand and Ravallion (1993) use slightly different nonlinear transformations of the relevant health indicator as the dependent variable in their analysis. While Anand and Ravallion (1993) do not offer much discussion about their choice of the transformation function, a major focus of Kakwani’s (1993) paper was on the derivation of the achievement index based on an axiomatic approach. The tests we performed so far indicated that this transformation seems to capture adequately the nonlinearity in the relationship. Kakwani (1993, p. 324, footnote 11) while referring to Anand and Ravallion’s (1993) paper noted that their transformation function is quite similar to his. We have also used Kakwani’s longevity data and an Anand and Ravallion (1993) type bivariate model and found that the estimated levels of longevity from the two models are very similar. Also, Li and Wang’s (1998) tests of the preceding kind used with this dependent variable yielded extremely similar results to those in Table 1. Given this and the slightly easier formula for Anand and Ravallion’s index, we chose their index rather than Kakwani’s for our analysis. So the infant mortality index in our study is given by the logarithm of the variable (IMR-5) where IMR denotes infant mortality rate and 5 refers to the “best” level of infant mortality.

Before going into regression analysis, first we present in Fig. 1 the scatter plot matrices of the variable involved in our regression analysis. The upper scatter plot is based on the untransformed variable (infant mortality rate, real GDP *per capita*, and health expenditure per head) while those in the lower scatter plot are based on the transformed variables (achievement index of infant mortality, welfare, and logarithm of health expenditure per head). The plots indicate some nonlinearity between
the index and the health expenditure variable, in both cases. We formally tested a linear relationship between the infant mortality index and the logarithm of GDP per capita against a purely nonparametric specification for this data set, using once again Li and Wang's test (1998), and found that we failed to reject the linear model at the 5% level of significance. We
also tested a linear relationship between infant mortality and GDP per capita and, in fact, rejected the linear model. The results from the tests are presented in the last six rows of Table 1. The results from these tests confirm our a priori belief that there exists a nonlinear relationship between the transformed variables. Given the scatter plots and this result, we are then going to assume a linear relationship between the infant mortality index and the welfare variable and are going to be concerned about possible nonlinearity in the public health expenditure variable.

Table 2 reports the regression results based on data on 26 countries in 3 different years from various specifications with infant mortality index as the dependent variable. The first four columns present results from various linear regression models while the fifth column presents the results from a semiparametric pooled model. From column one to column two the fit improves slightly as the logarithm of the variable public health expenditure per head, ln(hexp), is added to the model. The signs are according to our a priori expectation, namely, everything else remaining the same as welfare increases infant mortality goes down and similarly as public health expenditure increases infant mortality decreases, all else being equal. Both variables are statistically significantly different from zero at 5% level of significance. The fit improves further as the model in the second column is augmented by a squared term in order to capture possible nonlinearity coming in through the health expenditure variable. The squared term is strongly statistically significant. The next specification has a cubic term in the

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Regression Results for Infant Mortality Index (n = 26, t = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS(1)</td>
</tr>
<tr>
<td>Constant</td>
<td>10.542</td>
</tr>
<tr>
<td>Welfare</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Dummy 87</td>
<td>−0.171</td>
</tr>
<tr>
<td>Dummy 92</td>
<td>(0.033)</td>
</tr>
<tr>
<td>ln(hexp)</td>
<td>−0.142</td>
</tr>
<tr>
<td>ln(hexp)²</td>
<td>−0.147</td>
</tr>
<tr>
<td>ln(hexp)³</td>
<td>−0.045</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.613</td>
</tr>
</tbody>
</table>
health expenditure variable added to it, but now none of the terms involving the public health expenditure variable are statistically significant.*

The above specifications assumed that the welfare and the health expenditure variables are exogenous to the model; otherwise, the OLS estimator will be biased and inconsistent. There is a possibility that the welfare variable may be correlated with the error term since it may not be exogenous to the model, i.e., the infant mortality index is not only affected by the logarithm of income but the latter is also affected by the former. This is a possibility and so we need to test for the exogeneity of the welfare variable in our model. In order to perform the Hausman (1978) test to check for such correlation, we require instruments for the income variable. Following Pritchett and Summers (1996) and the discussion therein, we have chosen the logarithm of the ratio of the gross domestic public investment to the GDP of a country, and the logarithm of one plus the black market premium, as possible instruments for the logarithm of income. We performed the Hausman test using Wu's (1973) statistic, by including the predicted value from a regression of the logarithm of income on a constant and the instrument(s) in our linear or quadratic specification, described above, and then checking the statistical significance of the predicted value. In all the cases considered, the estimated coefficient on the predicted value was not statistically significant, and we failed to reject the null hypothesis of no correlation at any desired significance level. This provided some support for the assumption of exogeneity of the income variable in our regressions above.

Given that mis-specification of functional form can have an impact on our results, we next estimate a semiparametric (SP) model with ln(hexp) modeled as an unknown function. We first test the null hypothesis of a parametric model with welfare, ln(hexp), (ln(hexp))^2, dummy variable for 1987 (dummy87), and dummy variable for 1992 (dummy92) as the independent variables against a SP specification with ln(hexp) appearing

*The dummy variable for the year 1987 is not statistically significant in any of the OLS specifications. We have also estimated the same specifications excluding the dummy variable for 1987. The results for the estimated coefficients on the welfare and health expenditure variables remain almost identical to the results in Table 2.

†The black market premium is measured as (black market exchange rate/official exchange rate)−1. The data on black market premium (%) and gross domestic public investment as % of GDP are available in the Global Development Network Growth data base.

‡We calculated the predicted value by running regressions with each of the instruments separately as well as with both instruments together.
as an unknown function using the test by Li and Wang (1998). The results are given in Table 3 and they show that for all the bandwidths considered we can reject the null model at the 5% level of significance. This led us to choose the SP specification over the linear model augmented by a squared term as our preferred model.

Given that the data are panel in nature, we have also estimated a standard random effects (RE) model and a semiparametric random effects (SPRE) model (Li and Ullah, 1998). The estimates from the RE model are given in column 7 of Table 2 and show that the welfare variable continues to be statistically significant, and the result is quite similar to that of the pooled data case [OLS(3) in Table 2]. We carried out the Breusch and Pagan (1980) Lagrange multiplier test for RE and it strongly rejected the null hypothesis of the variance of the country-specific error term being zero. Given that the Hausman test (1978) failed to reject the null hypothesis of no correction between the country-specific error term and the regressors, we prefer the RE model over a fixed effects model. However, as in the pooled model case, the issue of mis-specification of functional form can still be a problem in this case. To address that issue, we estimated a SPRE model following Li and Ullah (1998).

The estimated nonlinear relationship between infant mortality index and ln(hexp) from the SP model along with its bootstrapped confidence interval* is presented in Fig. 2, while that from the SPRE model is presented in Fig. 3. By providing a confidence interval over the entire range of the logarithm of the health expenditure variable, we are able to offer a better representation of the nonlinearity in the relationship relative to the simple linear benchmark. The shape of the plot of the relationship between the index and the logarithm of the health expenditure from the SPRE model is very similar to that from the pooled SP model. The

| TABLE 3 | Tests of Quadratic Parametric Model Against Semiparametric Model |
|-----------------|-----------------|-----------------|-----------------|
| c = 0.8         | c = 1.0         | c = 1.2         | c = 1.4         |
| c.v. a          | −0.038          | −0.271          | −0.481          | −0.610          |

*a.c.v. is the bootstrapped 5% critical value based on 399 replications.

---

*We used the wild bootstrap to obtain the standard errors of the estimate and then used the normal approximation method to calculate the confidence bands. We chose the normal approximation method since the latter does not perform well when the sample size is small because of the importance this method attaches to the tails of the distribution (DiCicco and Romano, 1988).
Figure 2: Infant mortality index against health expenditure from a semiparametric pooled model.

Figure 3: Infant mortality index against health expenditure from a semiparametric random effects model.
estimated coefficient of the welfare variable is also very similar for the two models. These results indicate that the unobserved country-specific factors are not very important in affecting the infant mortality index since their inclusion did not change the results much. So given that it is relatively easier to calculate the pooled models and the results are similar across the different specifications, we prefer the pooled models in this case.

These results show that the public health expenditure has an effect on the standard of living as measured by the infant mortality index and that the relationship may not be linear as considered in the existing literature. However, the importance of income in influencing standard of living does not diminish much as the public health expenditure variable is included in the model. The results show strong and statistically significant effects of income on the infant mortality index even after we have taken into account the nonlinearity in the public health expenditure variable. This result is in contrast to Anand and Ravallion’s (1993) result that the effect of income on the standard of living diminishes substantially and may even become statistically insignificant once other variables like public health expenditure per head is added to the regression. It should be noted that Anand and Ravallion also had a poverty measure as an additional regressor in their models, involving life expectancy and literacy indices, but not in the model with the infant mortality index. The latter was due to the lack of data on the poverty measure for the case considered.

We faced a similar problem and were unable to obtain reliable data on the poverty measure for the 3 years without losing more countries from our sample. So we are unable to control for this particular variable. We recognize that this has the potential of biasing our estimate of the income coefficient. However, two things are to be noted in this respect. First, the inclusion of the poverty variable seems somewhat ad hoc in Anand and Ravallion’s (1993) paper, so it is not certain whether this is a relevant regressor or not. Second and more importantly, the purpose of this study was to analyze the effect of the inclusion of a variable representing public provision of health services on the estimated income coefficient. We found that the income variable continued to be highly statistically significant and that the inclusion of the health expenditure variable did not reduce its importance much,* contrary to what some researchers have suggested.†

*The coefficient estimate changes from 0.880 to −0.711 with the inclusion of the public health expenditure variable.
†It should be noted that the issues relating to possible nonstationarity of the data have not been taken into account in this study. However, none of the existing studies with which we are trying to compare our results have done so either. This issue remains an area of future research.
One other thing to note is that, like Anand and Ravallion (1993), we chose public health expenditure per head to represent public provision of health services. However, as Chakraborty (1999) discussed in her paper, what might be more important in developing countries are the coverage of public services and the nature of the orientation of the programs rather than the public health expenditure figure in itself. Also, private health expenditure seems to be relatively more important in developing countries as compared to developed countries. Unfortunately, reliable data on some of these variables are hard to obtain and hence a study involving these other variables remains an area of future research as better data become available.

4 CONCLUSION

We first tested for mis-specification in the functional forms of the linear models proposed by Kakwani (1993), who studied the effect of income on the achievement index and improvement index of life expectancy as an indicator of the standard of living of a country. Using a test by Li and Wang (1998), we found that the data supported the linear specifications as opposed to the nonparametric specifications put forth by Chakraborty (1999). This highlights, in general, the usefulness of performing a test such as the one proposed by Li and Wang before deciding on a nonparametric specification as the preferred model. The nonparametric estimator is biased in a finite sample and its attractiveness comes from its asymptotic properties. So given that we normally deal with relatively small-sized samples in cross-country studies, it is preferable to use a parametric model, if it is supported by the data, since the estimators will then have known finite sample properties and much faster speed of convergence.

We were also interested in the claim by certain researchers (Iseman, 1980; Sen, 1981; Anand and Ravallion, 1993; among others) that it is not economic growth but public provision of services which is more important in influencing the standard of living of a country, and that the strong and statistically significant effect of income vanishes once public provision of services has been accounted for. So we studied the effect of including another explanatory variable representing public provision of health services on the estimated coefficient of the income variable. The issue of nonlinearity in the relationship of this variable to the health index used was tested and we analyzed a semiparametric model that avoids the mis-specification of the functional form. Irrespective of the specification used, income continued to have a strong and statistically significant effect on the
health indicator. Further, controlling for unobserved country-specific
effects did not seem to impact the results much.

In general, the usefulness of the semiparametric model is that it
allows us to provide a pictorial representation of the nonlinearity in the
relation between the dependent variable and the variable* that appears as
an unknown function in the model, along with its confidence interval. The
methodology allows the data to decide the functional form and does not
impose any ad hoc functional form a priori, hence avoiding the issue of
functional form mis-specification. It is a very useful tool when one has a
number of regressors and the prior knowledge that some of them are
linearly related to the dependent variable. This kind of model avoids the
problem of the “curse of dimensionality,” which makes the purely
nonparametric estimators to be of little use if there are more than three or
four regressors in the model. On the computational side, these estimators
are not very burdensome either. Even though there are not many software
packages that would automatically perform a semiparametric estimation,
as long as those programs can do nonparametric estimation, they can be
quite easily manipulated to calculate semiparametric estimators as well.

ACKNOWLEDGMENTS

I thank the seminar participants at the University of Victoria and
an anonymous referee for their helpful comments on an earlier draft of
this chapter. I am very grateful to Thanasis Stengos for his help with some
of the GAUSS codes used for this study. The financial support from
the internal grant provided by the University of Victoria is gratefully
acknowledged.

REFERENCES

role of private incomes and public services. Journal of Economic Perspectives
7, 133–150.

*More than one variable can appear as the unknown function in the model, so to
obtain the individual effects one can follow the additive partially linear approach and
use marginal integration. See Section 2.1 for references.


Achievement index, 481, 486, 488
AIC (see Information criterion, Akaike’s)
Animation, 6
ARCH, 326–327, 346–347
Australian Bureau of Statistics, 43
Autocorrelation (see also Autoregressive process)
    positive, 44
    of residuals, 44, 326n
Autoregressive (AR) process
    (see also Autocorrelation), 38, 212, 213, 328, 360
    coefficient in, 327, 347
    exponential, 462
    filter using, 131
    functional coefficient, 455
    integrated moving average, 4
    model for, 4, 176
    moving average (ARMA), 333, 346
    parameter in, 329n, threshold (TAR), 356, 455, 461
Bartlett correction, 12
Bayes factor, 219, 220, 244, 250, 251
Bayes’ rule (see Bayes’ theorem)
Bayes’ theorem, 219, 266, 278
Bayesian inference
communication of results of, 251–258
computer-aided, 264
with conjugate prior distribution, 217–218
with hierarchical prior distribution, 220–222
intercept correction in, 389
and large data sets, 2
and model comparison, 218–220, 243–251
and model prior probabilities, 218–219
and model selection, 41
posterior density (distribution)
    tractable, 217
posterior odds ratio in, 219
and posterior simulation methods, 223–243
predictive density (distribution)
    for, 283–285
prior density (distribution) for, 2, 41, 212, 215, 218, 237, 250, 258, 385–386
conjugate, 216, 217–218, 227
construction of, 243
improper, 253
Bayesian inference
inequality-restricted, 274
Litterman, 386
noninformative, 253, 264, 266, 276
subjective, 253
tightness of, 390
two-stage hierarchical, 221
robustness of, 253
simulation methods for, 209–258
software for, 210, 237–240, 249–250, 254
Bias (see Estimator)
BIC (see Information criterion, Bayesian)
Bootstrap
block, 70, 82
conditional, 469
corrected, 71
coverage probability for, 69–90
critical values from, 484
method of, 25, 445
naive, 453, 458, 460, 463
parametric, 13, 21, 23n
recursive, 469
sieve, 82
test, 69–82
wild, 453, 458–460, 463, 487, 493n
Brownian motion, 98, 179
Bubble, price (see also Volatility, excess), 316–317, 318, 323–324, 327, 335, 344, 347
Calculus, fractional matrix, 11
CAPM model, 8
Causal ordering, 91
Causality
Granger, 91–145
long-run, 96
short-run, 96
Central Bank
European, 372
Norwegian, 387
Central limit theorem, 225, 235, 253
Chaos, 352, 361–362
Characteristic function, 73n
Cointegration (see also Data, cointegrated), 175, 391
and causality testing, 91–145
and stock prices, 323
Engle and Granger test for, 386
Johansen test for, 386
Collinearity (see also Data, ill-conditioned), 301
Computation (see Computing)
Computing
accuracy of, 248, 250
burden of, 245, 247, 325, 415, 418, 480, 497
difficulty with, 307
efficiency of,
excision time for, 241, 243, 250, 258, 335, 403, 463, 485
and FCM algorithm, 414–415
intensive nature of, 316
power of, 9
problems with, 165
rounding error, 308
speed of, 1, 9
Conditioning number (see also Data, ill-conditioned), 301
Cornish-Fisher expansion, 74, 75, 77
Cp (see Mallows’ Cp)
Cross validation (see Generalized cross validation)
Cumulant, 73, 74, 77, 78, 79
Curse of dimensionality, 408, 444
Data
achievement index, 485
animated (see Animation)
availability of, 1, 9
beef price, 440
Canadian earnings, 306–309, 419–422
cell phone, 200
chicken consumption, 440
chicken price, 440
cointegrated, 91–145
concrete price, 200
consumer price index, 372
consumption expenditure, 423
cross-sectional, 3, 7–8, 428, 471
dependent, 70, 72, 453
weakly, 452
difference, 94, 144
exchange rate, 359, 369–372, 434, 436
financial market, 317
gas, 200
Global Development Network (GDN), 486
gross domestic product (GDP) per capita, 485
ill-conditioned (see also Collinearity, Conditioning Number, and Multicollinearity), 291, 293, 306, 312
improvement index, 485
income inequality (see also Gini coefficient), 432
inflation rate, 372
integrated, 91–145
seasonally, 175–203
interest rate, 427
knitted fabric sales, 199
large sets of, 1–10
lot size, 240
missing, 6
money stock, 427
motorcycle, 301–304
nonannual, 175
non-stationary, 175, 312
orbit, 366, 368
output, 427
panel (see also Panel data), 2, 3
Panel Survey of Income Dynamics, 242
quality of, 6
real disposable personal income, 423
per capita, 440
real estate price, 240
retail trade, 43
reliability of, 495–496
scanner, 359
seasonal, 179
seasonally adjusted, 422–423, 434
semiannual, 175–203
share price (see Stocks)
stationary, 91, 176, 203
stock price (see Stocks)
tick-by-tick, 5
time-series (see Time-series)
unemployment rate, 372
wheat yield, 280

world tables, 43
Delta method, 78, 80, 246
Difference equation, 363
Differential equation, 223n
Distribution
asymptotic (see also Distribution, limit, non-standard), 179, 253, 347, 472, 484
chi square (see also Test, chi square), 25, 97–98, 131, 153, 254, 257
empirical, 74n
exact, 20
gamma, 224
inverted, 42
Gaussian (see also Distribution, normal), 7, 12, 15, 16, 17, 24
joint, 222
high frequency, 5
invariant, 231, 233, 234, 236
limit, non-standard (see also Distribution, asymptotic), 98, 180
lognormal, 334n
multinomial, 151
multivariate, 301
non-Gaussian, 12
non-standard, 95
normal (see also Distribution, Gaussian), 224, 250, 253, 254, 270, 285, 334n, 472
bivariate, 213
independent, 243
multivariate, 42, 154–155, 243, 265, 267, 275, 279, 290
truncated, 278–279, 286
standard, 484
univariate, 223
posterior (see Bayesian inference)
prior (see Bayesian inference)
reference, 295
t- (see also Test, t-), 163
multivariate, 268, 270, 279, 284
univariate, 270
two-sided, 75
univariate, 224, 295
Wishart, inverted, 267, 269, 279, 284, 285
Drift, 178, 179, 184, 185, 201, 386–387, 388, 403
Dummy variable, 389–391, 492
annual, 179, 180, 181, 184, 185, 201
Edgeworth expansion, 71–72, 74, 75, 77, 78, 80, 460
Equivalence scale, 282
Equity premium puzzle, 343
Error correction model, 96
vector (VEC), 102, 105–107, 129, 145, 386–387, 390–391, 403
Estimating function, 294
Estimation (see Estimator)
Estimator (see also Regression)
asymptotic properties of, 456
Bayes (see Bayesian inference)
based, 150, 154, 157, 159, 160, 297, 311, 316, 322, 329n, 492
canonical modulation, 299
consistent, 2, 152, 236, 296, 386
weakly, 418, 445
efficiency of, 96, 297
efficient, asymptotically, 152
extremum, 297
fuzzy (see Fuzzy econometric modeling)
inconsistent, 480, 492
inefficient, 322, 480
kernel (see Estimator, nonparametric, and Regression)
least absolute deviations, 67, 415
least squares (LS)
generalized, 25–26, 31, 267, 390
indirect, 389
local linear (LLLS), 454–455
trimmed, 415
M-, 415
maximum empirical likelihood (EL), 149–166
maximum likelihood (MLE), 13, 16, 24, 25–26, 40, 106, 243, 258, 267, 274, 292, 312
quasi-, 26
method of moments (MOM), 150
method of regularization (MOR), 297, 312–313
minimax, 292
mixed regression, 386, 387, 484
Nadaraya-Watson, 311, 459
normal, asymptotically, 152, 296
panel (see Panel data)
parametric, 407
precision, 300, 301
principal components (PC), 300, 302–304, 359
quantile, 4
ridge, 297, 299, 300, 302, 304
iterative, 297
robust, 415
semiparametric, 482–483, 485, 492
single-equation, 280
Stein, 292, 293, 313
modified, 293
positive-rule, 292–293
superconsistent, 388
SUR (see Seemingly unrelated regressions)
three stage least squares (3SLS), 389
iterated, 389
two-stage least squares (2SLS), 150, 154, 156, 158–166, 388, 390
two-step, 267
unbiased, 294, 298, 299, 308, 312
European Monetary Union, 372
Exogeneity, 215
Federal Reserve Board, 315–316
Filter
autoregressive (see Autoregressive process, filter using)
Hodrick-Prescott (HP)
(see Time-series, trend in)
Index

[Hodrick-Prescott (HP)]
 moving average, 103
 stationarity, 175
Forecasting
 Bayesian, 266, 283–285
 conditional distribution for, 4
 ex ante, 92, 101–103, 107–108
 out-of-sample, 5, 92, 102
 real-time, 102, 385–404
Fuzzy clustering, 407–446
 c-means (FCM) algorithm for,
 412–415, 417, 419–420, 423,
 442, 444, 472
 binary splitting (BS) variant of, 414
 binary splitting direct search
 variant of, 414
 KKZ variant of, 415
 mountain method variant of, 415
 simple cluster-seeking (SCS)
 variant of, 414
 subtractive clustering variant
 of, 415
Fuzzy econometric
 modeling, 415–418
Fuzzy logic, 410–412, 444
Fuzzy set, 410–412, 444
 degree of membership for, 410, 411,
 412, 413, 444
 membership function for, 420, 425,
 427, 432, 436, 438
Gamma function, 270
General to simple methodology, 7
Generalized cross validation (GCV), 38
Gibbs sampler (see Markov-chain
 Monte Carlo)
Gini coefficient (see also, Data,
 income inequality), 428
Granger causality (see Causality)
Hadamard product, 295
Harmonic mean, 246–249
Hodrick-Prescott (HP) filter
 (see Time-series, trend in)
Hyperparameter, 221
Improvement index, 481, 486
Income inequality (see also Gini
 coefficient and Data, income
 inequality), 428
Indicator function, 213n, 458
Indirect least squares (see Estimator,
 least squares)
Inference
 Bayesian (see Bayesian inference)
 ill-posed, 291, 312
 Information (see Kullback-Leibler
 information)
Information criterion
 Akaike’s (AIC), 38, 40, 44, 101,
 107–108, 130, 143–145
 Bayesian (BIC), 38, 40, 44, 67, 326,
 333
 Hannan and Quinn’s (HQ), 38,
 40, 44
 Schwartz’s (SIC), 92, 101, 107–108,
 129–130, 143–145
Input-output relationship, 416
Instrument (see also Instrumental
 variable), 155, 492
Instrumental variable (IV) (see also
 Instrument), 149, 151, 156, 164
Integration
 analytical, 249, 267
 bivariate, 223
 marginal, 483
 Monte Carlo, 325,
 numerical, 284, 320, 325, 330, 335
 by simulation, 222
 Intercept correction, 387, 388–391
Inverse problem, 294, 297
Kolmogorov’s representation
 theorem, 356
 Kullback-Leibler (KL)
 information, 295
Kurtosis, excess, 338, 343
Kuznets’ U-curve, 428–432
 dynamic effects and, 432
 nonstationarity data and, 432
Lagrange multiplier test (see Test)
Latent variable, 220–222
Law of the excluded middle, 411
Least squares (see Estimator, least
 squares)
Likelihood function, 211, 221, 251,
 254, 265, 276, 278

Copyright © Marcel Dekker, Inc. All rights reserved.
MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016
[Likelihood function]
marginal, 219, 220, 237, 243–244, 245–246, 248–249, 250, 251, 266
for probit model, 251
unbounded, 212
Likelihood ratio test (see Test, likelihood ratio)
Lindeberg-Levy central limit theorem (see Central limit theorem)
Logistic function, 354–355, 357
Loss
balanced, 293, 298, 299
squared error (SEL), 154, 157, 292, 297, 298
squared error prediction (SEPL), 299
Lyapunov exponent, 352, 361–364, 366, 368–371, 381
M-estimator (see Estimator)
Mallows' Cp, 308, 311
Marginal propensity to consume (m.p.c.), 425
Markov model, 319–320, 332
Markov-chain Monte Carlo (MCMC)
and Bayesian inference, 263
burn-in iterations for, 240, 273
central limit theorem for, 236
corvergence of, 234–235, 273
two-block, 231
Hastings-Metropolis algorithm for, 224, 229–232, 244, 264, 274–277, 279, 281, 283
corvergence of, 235
methods for, 224, 263–264, 273–274, 286
Metropolis-within-Gibbs algorithm for, 231–232
numerical accuracy of, 235–237
Maximum empirical likelihood (see Estimator, maximum empirical likelihood)
Maximum likelihood (see Estimator, maximum likelihood)
Mean absolute error (MAE), 422, 423, 427, 432, 441
of forecast, 102
Mean squared error of estimation, 157
of forecast, 102, 393, 396–397, 463
of prediction, 157
root (RMSE), 422, 423, 427, 432, 441
Method of moments (see Estimator, method of moments)
Missing value, 6
MLE (see Estimator)
Model comparison
Bayesian approach to, 218–220, 243–251
Model selection
backward, 308
Bayesian approach to, 41
for causal orderings, 91
general to specific (see General to simple methodology)
for linear regressions, 37–67
for VAR models, 92, 100, 145
MOM (see Estimator, method of moments)
Moment conditions, 149, 150, 153
Monte Carlo experiment, 24–30, 33, 39, 43–67, 71, 74n, 92, 94, 103, 153, 154, 185, 325n, 327, 328, 335–336, 403, 487
integration (see Integration, Monte Carlo)
method, 45, 300, 306, 317, 324, 458
p-value, 21, 23
procedure (see Monte Carlo, method)
sampling, 149–150, 154, 165
simulation, 41, 184, 325, 333–335
error of, 325n
test, 11–33
variance reduction in, 330n
Moving average (MA)
in errors, 108, 14–144, 181
model for, 2, 326
process for, 38, 92, 105, 130, 145, 328
Multicollinearity (see Collinearity, and Data, ill-conditioned)
Index

Neyman factorization criterion, 217n
Neural network, 351–382, 472
   activation function in, 353–354
   artificial, 352
   layers of, 353, 364–365, 373, 376, 378
   node reduction in, 358
   pruned, 360, 369–370
   pruning methods for, 357–358
   reduced (see Neural network pruned)
   three-layer feed forward, 352, 356
   two-stage, 352
   weight reduction in, 358
Noncausality (see Causality)
Nuisance parameter, 13, 18, 97
OLS (see Estimator, least squares)
Optimization
difficulty with, 45, 403
numerical, 45, 351
Ordinary least squares (see Estimator, least squares)
Outlier, 4, 6, 415
Overfitting, 312, 357
P value
   local Monte Carlo (LMC), 21
   maximized Monte Carlo (MMC), 21
   simulated, 23
Panel data (see also Data, panel)
   balanced, 486
   estimation with, 2, 3, 6, 479–497
Petroleum Monitoring Agency Canada, 200
Phillips curve, 372–381
Pivotal statistic (see Statistic, pivotal)
Posterior odds ratio (see Bayesian inference)
Power (see Test, power of)
Prediction (see Forecasting)
   Preliminary test, 387
   bias from, 38
Pricing model
   capital asset (see CAPM)
   fundamental, 318–322
   Donaldson and Kamstra (DK), 320–322, 323, 332–333, 346
   Gordon, 319, 332, 346
Markov, 319–320, 332
Prior (see Bayesian inference)
Probit model, 242–243, 251, 286
Purchasing power parity (PPP), 371, 485
Quadrature (see Integration)
Random walk, 178, 185, 385
Regression (see also Estimator)
   auxiliary, 184, 185, 459
   kernel, 311, 432, 444, 483
   multivariate, 407
   linear (see also Estimator, least squares), 38, 39, 43, 215, 240, 310, 483, 491
   multivariate linear (MLR), 12, 18
   nonparametric, 310, 408, 427, 438, 490
   parametric, 451
   polynomial, 307, 310
   ridge (see Estimator, ridge)
   semiparametric, 452
   with principal components, 303
   spline (see Spline)
   varying coefficient, 456
Reversibility condition, 230
Ridge estimator (see Estimator, ridge)
Risk, 155, 156, 158, 161, 290, 292, 297, 304, 313
   empirical, 303
   Risk premium, 322n
Sampling
   acceptance, 224–225
   importance, 224, 226–227, 238, 244
   Seemingly unrelated regressions (SUR, SURE)
      Bayesian inference for, 263–287
      inequality restrictions in, 278–279
      nonlinear, 275–276
      restrictions on parameters of, 11–33, 277–279
      testing restrictions in, 11–33
      unbalanced, 446
   Serial correlation (see also Autoregressive process, and Moving Average), 238, 316n, 326
Serial dependence (see Serial correlation)
Shares (see Stocks)
SIC (see Information criterion, Schwartz’s)
Simulated annealing (SA), 45, 50, 52, 66
Simulation (see also Monte Carlo) of economies, 315–348
experiment (see Monte Carlo, experiment)
Simultaneous equations model, 286
Size (see Test, size of)
Skew, 338, 343
Spline, 310, 311, 444, 472
Standard of living, 479–497
Statistic, pivotal, 12, 18, 23
sufficient (see Sufficient statistic)
Statistics Canada, 200
Statistics New Zealand, 199
Stein estimator (see Estimator)
Stocks
data for, 359
market for, 9
on New York Stock Exchange, 8
prices of, 315–348
ex post rational, 322, 332
Structural break, 4
in DVAR model, 390, 392–393
in MDVAR model, 389
testing for, 3, 387
and seasonal unit root tests, 203
Structural equations, 149–166
Sufficient statistic, 217, 218
SUR (see Seemingly unrelated regressions)
SURE (see Seemingly unrelated regressions)
Taylor’s theorem, 75n
approximation, 455
series expansion from, 76, 81
Test
Anderson-Darling, 326n
asymptotic, 26n, 33, 460
bias of, 29, 181
bootstrap, 69–82
nonparametric, 451–472
bounds, 13
Monte Carlo (BMC), 12
chi-square (see also Distribution, chi square), 26, 29, 31, 129, 240
for bubbles, 324
Camerer, 323, 346
Campbell-Shiller, 323, 346
conditional moment, 459
nonparametric, 472
consistent, 452
Cramer-von Mises, 326n
Diebold-Mariano, 102
empirical likelihood ratio (ELR), 162
Engle and Granger, 386
ex ante, 129
F-, 11, 26, 29, 106, 129, 201n, 452
nonstandard, 177, 201n
feasibility of, 33
based on generalized method of moments (GMM), 72
Hausman, 492
Wu’s version of, 492
Johansen, 386
Kolmogorov-Smirnov, 326n
L, 459–460, 461
Lagrange multiplier (LM), 12, 162
with large samples, 2
Li and Wang, 482–483, 485, 489
likelihood ratio (LR), 12–13, 16, 18–19, 21, 452
quasi- (QLR), 25, 31
Monte Carlo, 12
bounds, 22
local (LMC), 26–27, 29
MRS1, 324
MRS2, 324
nonparametric, 452, 472
pairwise, 37
parametric, 451
preliminary (see Preliminary test)
power of, 26, 29, 33, 181, 185, 461–462, 465, 471
Q, 459, 461
score (Rao’s), 458
Shapiro-Wilks, 326n
simulation-based, 11–33
size of, 26, 27, 33, 95, 129, 143, 181, 185, 461–462, 465
specification, 451–472, 480, 482, 483–484
Index

[Q]
- (see also Distribution, t-), 179, 181, 201n
  nonstandard, 177, 201n
T, 457, 460, 461
  S version of, 458, 460
  R version of, 458, 460
  unit root, 4, 371
  Dickey-Fuller (DF), 176, 181, 203, 323, 346
  augmented (ADF), 178, 199, 201
  F-, 179, 201n
t-, 179, 181, 201n
  Hylleberg-Engle-Granger-Yoo (HEGY), 176–178, 181, 203
  with monthly data, 176
  seasonal, 175–203
  with semi-annual data, 175–203
  with quarterly data, 176
Wald
with empirical likelihood, 153
  for Granger noncausality, 94–99, 101, 105–107, 129, 131
  sequential, 96–97, 143
in SUR model, 12, 24–26, 31
with two-stage least squares, 163
Three stage least squares (see Estimator, three stage least squares)
3SLS (see Estimator, three stage least squares)
Time-series, 469
  artificial, 438
  estimation with, 8–9
long, 3
  mixing conditions for, 236
  multiplicative model for, 434
  nonlinear, 452
  seasonal pattern in, 434
  seasonally integrated, 175–203
  trend in, 178, 179, 184, 185, 306
  cubic, 278
  extraction of, 434–439
  Hodrick-Prescott (HP) filter for, 435–436, 438, 444
Tobit model, 285, 286
Trend (see Time-series)
  Trimming parameter, 291–313
TSSS methodology (see Fuzzy econometric modeling)
Two-stage least squares
  (see Estimator, two stage least squares)
2SLS (see Estimator, two stage least squares)
U-curve (see Kuznets’ U-curve)
Unit root (see Data, integrated)
  test (see Test, unit root)
VAR (see Vector autoregression model)
VEC (see Error correction model, vector)
Vector autoregression (VAR)
  Bayesian (BVAR), 386, 390
  in differences (DVAR), 387, 390–393, 396–397
  Bayesian, 396–397
  mixed drift (MDVAR), 389–391, 393, 396–397, 403
  Bayesian, 396–397
Volatility, excess
  (see also Bubble), 316–317, 323–324, 336
  Wald test (see Test, Wald)
Walds phenomenon, 457
World Bank, 43